Genomic insights into bacterial adaptation during infection

Genomic insights into bacterial adaptation during infection

A dissertation presented

by

Tami Danielle Lieberman

to

The Committee on Higher Degrees in Systems Biology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Systems Biology

Harvard University

Cambridge, Massachusetts

January 2014

Genomic insights into bacterial adaptation during infection

Abstract

Bacteria evolve during the colonization of human hosts, yet little is known about the selective pressures and evolutionary forces that shape this evolution. Illumination of these processes may inspire new therapeutic directions for combating bacterial infections and promoting healthy bacteria-host interactions. The advent of high-throughput sequencing has enabled the identification of mutations that occur within the human host, and various tools from computational and evolutionary biology can aid in creating biological understanding from these mutations. Chapter 1 describes recent progress in understanding within-patient bacterial adaption, focusing on insights made from genomic studies.

Chapters 2 and 3 investigate how the opportunistic pathogen *Burkholderia dolosa* evolves during long term infections of people with cystic fibrosis, studying patients infected during the same outbreak in Boston. Chapter 2 reports the genomic sequencing of 112 isolates taken from 14 patients over a period a 16 years. Phylogenetic reconstruction identifies a likely transmission network between patients and reveals multiple lung-to-blood transmissions during disease progression. Seventeen genes underwent parallel evolution in these patients, revealing new genes important to bacterial survival *in vivo* and highlighting the role of a particular oxygen-dependent gene regulation pathway.

Chapter 3 studies the co-existing *B. dolosa* diversity within single sputum samples taken from each patient, using both colony re-sequencing and a population deep sequencing approach. The intraspecies diversity within each sample is vast and suggests that the diverging lineages co-

exist for many years within patients. Furthermore, these diverging lineages evolve in under the pressure of selection and in parallel, enabling the identification of genes undergoing selection from a single clinical sample.

Chapter 4 describes two ongoing collaborations that extend these approaches to other infections, directly addressing the spatial component of bacterial diversification. Chapter 5 discusses future potential directions and consequences of the further study of bacterial evolution within the human body.

Table of Contents

Acknowledgements

I feel incredibly fortunate and honored to have reached this stage in my life and career. In addition to the acknowledgements specific to each Chapter, I would like to profusely thank many people for their contributions, motivation, and support, including:

My thesis advisor, Roy Kishony, for his scientific insight and constant support, for motivation when necessary, for providing an excellent example of patience and respect for others, and for giving me so many opportunities for learning and development;

Past and present members of the Kishony Lab, who together with Roy have created the most amazing place to work and learn: Laura Stone for being a constant source of excellent advice; Adam Palmer, Daniel Schultz, and Ylaine Gerardin for thought-provoking discussions and being superb baymates; Michael Baym and Remy Chait for teaching me the power of building; Hattie Chung for joining me in the quest to understand bacterial within-patient evolution; Erdal Toprak for making weekends in the lab seem normal; Seungsoo Kim for being an incredibly skilled undergraduate collaborator; and Kalin Vestigian, Eric Kelsic, Ilan Wapinski, Joel Yao, Justin Meyer, Adrian Grenada, Joy Jiao, Kathy Wang, and Morten Ernebjerg for sharing their expertise, support, energy, and camaraderie;

My co-author Jean-Baptiste Michel for introducing me to the wonders of within-patient evolution and for teaching me how to be an effective graduate student, collaborator, and computational biologist;

My collaborators Greg Priebe, Alex McAdam, Kelly Flett and Sara Vargas for enabling these studies, lending their expertise, and patiently teaching me about clinical research; Ted Cohen and Doug Wilson for including me in their ongoing study and for teaching me about tuberculosis, the realities of poverty, and dedication to a cause; Nick Leiby for thousands of data points; Will Harcombe for expert mentorship of my first rotation; Christina Agapakis and Pam Silver for expanding my in interest in our relationships with bacteria; and Jake Wintermute for teaching me about flux balance analysis and how to give a chalk talk;

The International *Burkholderia Cepacia* Working Group for accepting me into the community and ensuring that discovery does not stop at lists of candidate genes; The Microbial Evolution Group and Microbial Sciences Initiative at Harvard for insightful discussions;

Chris Marx for teaching me so much about microbial evolution during OEB 192, my rotation in his lab, and his service on my advisory committee; Chris Burge for providing me with the tools for computational biology in 7.91; Uri Alon for imparting the wisdom of the cloud so early on in my scientific career; my professors at Northwestern University who provided a rigorous background in biology and mathematics; members of the Widom Lab who first taught me that scientists are excellent people; and especially the late Jon Widom for his support, humanity, and inspiring clarity;

(continued on next page)

Curtis Huttenhower, Marc Lipsitch, and Steve Lory for their helpful advice and healthy criticism as past committee members;

Sam Reed, Becky Ward, Tenzin Phulchung, and the rest of the Systems Biology support staff for enabling us to focus on the science; Sarah Boswell and members of the Springer Lab for assistance with the MiSeq;

My exam committee, Sarah Fortune, Michael Laub, Pardis Sabeti, for their comments on this work; and especially Tim Mitchison for his service on many committees and his enthusiastic creativity and support;

My dear friends Anna Turetsky, Catalina Romero, Josh Reyes, Mashaal Sohail, Greg Koytiger, and Paul Cowgill for their wise words and for making Boston a warm and lively home; Dorit Goikhman and Katie Funkhouser for their excellent support and laughter from a distance and for motivating complaints of law school; Tara Martin, Ashley Wolf, Angela Depace, Galit Lahav, and the rest of the Systems Biology Women's group for their open arms; and the Boston Ultimate Disc Alliance community for their friendship and the opportunity to run around;

And, of course, my family for their bountiful love and laughter: my parents, Sherry and Mike Lieberman, for their constant support, for their inspirational consistency and excellence in work and life, and for accepting that this is not an MD; my twin, Alex, for sharing his healthy skepticism—which we could all use more of; and my older brother, Steve, for teaching me at a young age not to fear the command line.

Chapter 1:

Genomic insights into bacterial adaptation and diversification within the human host (a review of the field)

Bacteria evolve during the colonization and infection of human hosts, adapting to the host environment, immune defenses, and antimicrobial therapy. Identification of the selective forces shaping this evolution can inspire new therapeutic directions for combating bacterial infections and promoting healthy bacteria-host interactions. Furthermore, an understanding of how bacteria diversify as they adapt can illuminate the molecular mechanisms of pathogenesis and is crucial for tracking transmission networks. Breakthroughs in DNA sequencing have recently enabled the identification of mutations occurring during infection, but parsing biological understanding from these mutations can be challenging. During my time in graduate school, significant strides have been made by us and others towards understanding genomic bacterial evolution within humans. This Chapter reviews approaches and challenges to understanding within-patient bacterial evolution, recent findings, and their implications.

Motivation

Bacteria evolve rapidly to meet new challenges; in the context of pathogens this poses a therapeutic challenge and a threat to human health. The evolution of bacterial lineages into important pathogens is a significant global health problem; approximately 160 newly emerging infectious diseases caused by bacteria have been discovered in the past 70 years[1]. Likewise, the rapid spread of drug-resistance in established pathogens poses a significant and mounting danger to human health[2,3].

Yet, the potent weapon of genomic evolution holds enormous potential for research. By tracking genomic changes that enable a pathogen lineage to thrive within our bodies, we can identify its weaknesses and design therapeutic strategies to heighten these challenges. Additionally, within-patient bacterial evolution generates diverse lineages that are separated by relatively few mutations, enabling straightforward genome-wide association studies for identifying the genomic basis of phenotypic changes[4,5]. Furthermore, we can use these genomic breadcrumbs to track the progression and transmission of pathogens throughout the body body[6-8].

With the speed and cost of genomic sequencing already so favorable (a few days and less than $30 per genome, including sample preparation) and getting better every year, the potential for rapid insight is enormous. The field of genomic within-patient bacterial evolution has been slow to grow, with studies of within-patient cancer evolution surprisingly leading the way in some cases despite the higher depth of sequencing required[9,10]. The investigation of within patient-evolution has greatly accelerated in recent years; since 2011, over 25 studies across a variety of pathogens have compared multiple genomes of the same species from the same patient, compared to 7 previously existing studies. The genomic progress towards understanding within-patient evolution of *Pseudomonas aeruginosa*[11] and *Staphylococcus aureus*[12] are

described in recent review articles. Here, I focus on the common themes emerging from the study of various pathogens—including parallel evolution and diversification—, common challenges to studying within-patient evolution across microbes, and the potential of large-scale genomic studies for illuminating epidemiology, clinical practice, and basic biology.

Bacterial adaptations during infection

Any discussion of within-patient evolution must begin with a discussion of cystic fibrosis (CF), as so much of what is known about bacterial evolution within the host comes from chronic infections of people with this common autosomal recessive disorder. People with CF lack a functioning copy of the CFTR chloride channel and have a number of impairments, including viscous mucous that is hard to clear and provides a habitat for long-term bacterial infections. Infections of people with CF are polymicorbial[13,14], but most people with CF become colonized by a single dominant strain which persists for decades, providing many opportunities for mutation and selection[15].

Adaptive evolution was noticed in *P. aeruginosa*, the most common CF pathogen, as early as the 1960's. This adaptive evolution was easy to spot—in many patients, colonies cultured in a later stage of the infection produced excess alginate, resulting in a very noticeable "mucoid" phenotype, whereas *P. aeruginosa* isolates from earlier on in infection or the environment do so at a much lower rate[16,17]. This change was shown to be genotypic rather than phenotypic[17], suggesting parallel evolution. Parallel evolution is a strong signature of positive selection and remains the dominant marker of within-patient adaptive evolution today.

Another reason for the emergence of CF infections as a model system is that, unlike many other infections in which antibiotic resistance is conferred via acquisition of a horizontally

transferred resistance cassette, these infections primarily acquire resistance through *de novo* mutation. As these infections are often acquired from environmental sources, they commonly start out as antibiotic sensitive and later, in a step-wise fashion, acquire the ability to tolerate multiple drugs[18]. For the same reason, there is now a rapidly growing body of work focusing on *Mycobacterium tuberculosis* within-patient evolution.

Since the 1960's, a suite of common phenotypic changes have been observed in *P. aeruginosa* during colonization of the CF lung, where later isolates exhibit particular traits more often than earlier isolates. Beyond antibiotic resistance and mucoid colonies, other traits associated with *P. aeruginosa* adaptation to the CF lung include changes to the cell envelope, loss of motility, and loss of quorum sensing[11]. The parallel evolution of these traits in many patients suggests that they are adaptive. Many studies have determined the genetic basis and biological roles of these phenotypic changes[19,20]; for example, mucoid phenotypes are associated with worse patient outcome[17].

Yet, the story is complicated because genetic investigations have shown that the same commonly mutated master regulators control many of these traits[11]. As a consequence, many heavily studied phenotypic transitions may not be directly selected upon. For example, while many studies have attempted to explain the selective pressure driving alginate production[17], it is possible that alginate production is a pleiotropic effect of selection on a different component of bacterial survival in the lung, such as cell wall stress[11]. Thus, reliance on phenotypic observations alone can bias the researcher towards changes that are not directly selected upon and against adaptive changes that are not readily visible. In contrast, genomic approaches offer an unbiased and rapid approach to identifying the drivers of adaptation within the host.

Whole-genome sequencing reveals genes under parallel evolution

With the advances in genome sequencing in the late 2000's came the ability to identify changes during infection on a genome-wide scale, in an unbiased manner. In a 2006 landmark study by Smith *et al*., the whole genomes of two *P. aeruginosa* isolates taken from the same individual, separated by 8 years, were sequenced using shotgun Sanger sequencing[21]. The authors identified 68 mutations separating these isolates, the majority of which caused amino acid substitutions or frameshifts. To identify which of these mutations were likely driving the adaptation of these infections, they screened pairs of isolates from dozens of other patients for mutations in 24 of the candidate genes using targeted Sanger sequencing. This study revealed that many of the mutations found were not common across patients, but that some genes were mutated in nearly half of the patients' later isolates. These commonly mutated genes, thereby implicated as some of the most significant contributors to *P. aeruginosa* survival in the body, include multidrug efflux genes and a quorum-sensing regulator.

This approach of comparing the whole genomes of a few isolates, separated by long periods of time, from the same individual has been emulated in many studies of various bacteria. Multiple *P. aeruginosa* infections have been tracked using genomic, transcriptomic, and proteomic approaches, and both similarities and differences from the genomic trajectory observed by Smith *et al*. have been reported[11,19,22,23]. A study of *Burkholderia pseudomallei* evolution during long term asymptomatic carriage of one patient demonstrated genome reduction in the host enviornment[24] and other changes thought to be towards commensalism[25]. In a unique study, Zdziarski *et al*. tracked evolution of an asymptomatic *E. coli* therapeutically introduced into the bladder of 6 individuals, finding parallel evolution across patients in multiple genes, including stress and virulence associated genes. Interestingly, some genes were mutated during

multiple independent introductions of the therapeutic strain into the same individual, but not in other patients, suggesting that adaptation can be host-specific[26].

Since 2011, the dropping cost of whole-genome sequencing has made a new approach possible: the simultaneous sequencing of dozens to hundreds of isolates from each species, spanning many individuals[6,27-32]. This large-scale approach enables systematic and straightforward diagnosis of adaptive parallel evolution across patients. Particular progress has been made in studies that have focused on sets of patients infected with very similar strains, simplifying both the identification of mutations (by use of alignment to a single reference genome) and interpretation of the mutations found[6,28,32]. These studies have use standard phylogenetic methods to identify which nucleotides are shared by these closely related isolates because of common ancestry, and which are due to parallel nucleotide evolution[6,28,32].

The first such study compared the genomes of 112 isolates of *Burkholderia dolosa* from an outbreak among patients with cystic fibrosis, spanning 14 patients during 16 years[6] (Chapter 2). This study identified 21 genes that were mutated multiple times during the outbreak and 17 genes mutated three or more times. Importantly, the mutation of a gene twice during these infections was consistent with a neutral model, and thus only genes mutated three or more times were likely adaptive. Consistent with this analysis, genes mutated three or more times had a strong enrichment for nonsynonymous mutations (measured by the canonical signature for selection, dN/dS, the relative rate of nonsynonymous over synonymous mutations), while genes mutated twice or once did not. This approach has also been used by Marvig *et al*. on an transmissible lineage of *P. aeruginosa*, with surprisingly similar results. More mutations were found in this lineage, requiring a higher threshold for parallel evolution of more than 6 mutations per gene. Despite this high threshold, they found many genes under parallel evolution. In both of

these studies, known antibiotic-resistance determinants and virulence factors evolved in parallel, but so did many novel genes, including several two-component systems in *P. aeruginosa* and an oxygen-dependent gene regulation system in *B. dolosa*. The inferred importance of these genes to bacterial survival suggests that their further investigation will be important to understand the drivers of within-patient evolution and might uncover new therapeutic directions. A third study by Golubchik *et al*. attempted to identify parallel evolution during asymptomatic carriage of *S. aureus,* but did not find any such evidence during across 13 healthy people carrying the same clonal complex[28]. This is consistent with other signals for purifying selection, and may reflect the different lifestyle or evolutionary history of these bacteria.

The ability to detect parallel evolution has turned out to be a singularly powerful tool for identifying adaptive evolution[6,32,33], as other signals for targets of selection are not applicable given the low number of mutations. This relatively small number of mutations found during infection does not provide a strong enough signal for the use of dNdS at the gene level. And while a strong positive or negative dNdS has confirmed that certain sets of genes are undergoing positive selection[6,32], a neutral dNdS should not be taken as evidence of genomic drift. For example, in the *B. dolosa* study mentioned, positive selection on some genes and purifying selection on other genes average across the genome to provide neutral signals of dNdS, which is in stark contrast to the strong evidence for adaptive evolution [6] (Chapter 2).

Clock-like evolution of bacterial infections

These genomic studies have revealed that single nucleotide mutations accumulate within patients according to a molecular clock[6,26,31,32,34], even as the rate of observed phenotypic change slows[27]. The finding that bacterial lineages accumulate mutations linearly over time despite

varying strength of selective pressures over time is perhaps surprising, but also echoes the findings in laboratory evolution experiments[35]. The rate of this molecular clock has been estimated for various strains using deliberate infections and subsequent tracking[26,34], paired clinical isolates[30,31], and larger collections of isolates taken from the same patient or group of patients[6,7,36]. These studies have provided estimates of the molecular clock for different species and lineages, with most being in the range of 0.5-10 per genome per year. These rates are similar to those from studies of outbreaks and global transmissions[37-39].

Various approaches can be employed to estimate the rate of the molecular clock, but the most common approach is to calculate the slope of the line of the number of mutational events versus time of collection. When doing this calculation, it is important that the number of mutational events is estimated relative to the most recent common ancestor or outgroup; because most within-patient populations diversify as they adapt (more on this later), using the number of mutations separating two isolates can overestimate the molecular clock.

Another important consideration is the role of recombination in supplying new alleles. If horizontally acquired genomic segments are treated as *de novo* mutations, the rate of evolution can be vastly overestimated—and the inferred phylogenies will be incorrect. The importance of recombination within patients varies widely between organisms, with some organisms, such as *Mycobacterium tuberculosis,* demonstrating no evidence of recombination despite ample study[40] and others, such as *Heliobacter pylori* and *Streptococcus pneumoniae*, showing high rates of recombination[41-43]. Comparison of highly sexual strains requires more complex analysis, as each isolate genome may have to be assembled separately *de novo,* whole-genome alignment must be performed, and recombination events need to be identified either before or during generation of a phylogenetic tree[39,42,44].

Hypermutation is common during infection

The rate of the molecular clock can sometimes change within a strain. One common trend in within-patient evolution is the emergence of hypermutation phenotypes, with increased mutation rates, during infection. Hypermutation is usually caused by a defect in DNA repair; these defects in DNA repair hitchhike along for the ride with beneficial mutations directly selected upon, which occur more frequently in these backgrounds[45,46]. Hypermutation has been observed frequently in bacteria colonizing the CF lung, including *P. aeruginosa*[47], *Haemophilus influenzae*[48], and *S. aureues*[49] and has also been observed in other pathogens, include *E. coli* infecting the urinary tract[46,50].

Recently, genomic studies have been able to detect hypermutation from sequence alone. In the described study by Morvig *et al*., a number of *P. aeruginosa* isolates from the DK2 lineage had many more mutations than would be predicted based on the sampling time and molecular-clock estimated from other isolates in the DK2 lineage. Further inspection revealed that only these isolates with excess mutations had defects in DNA repair genes, with each isolate's spectrum of mutations matching the known role of these genes in DNA repair[32]. Similarly, in our second study of multiple patients infected during the same *B. dolosa* outbreak, we observed that one patient's isolates had an excess of transition mutations (purine to purine and pyrimidine to pyrimidine) and a corresponding mutation in mismatch repair[29] (Chapter 3).

Genomic paths to *in vivo* antibiotic resistance

A complete understanding of antibiotic resistance loci will enable rapid whole-genome sequencing for the diagnosis of antibiotic susceptibilities[51] and will provide a deeper

understanding that may one day lead to therapies which will more effectively prevent the spread of resistance. Even for *M. tuberculosis,* an infection in which the genetic causes of antibiotic resistance mutations are heavily studied and for which genotypic assays are already clinically used to make decision on antimicrobial therapy, studies continue to find new mutations which confer resistance to widely-used antibiotics[4].

Comparing the whole-genomes of sensitive and resistance isolates taken from the same individual can rapidly identify *de novo* mutations responsible for resistance. A landmark 2007 study by Mwangi *et al*., identified mutations responsible for the sequential emergence of resistance to multiple antibiotics during a persistent bloodstream infection with *Staphylococcus aureus*[52]. Since then, many studies of paired isolates from acute infections have used whole genome sequencing to rapidly identify known and novel antibiotic resistance determinants in *Acinetobacter baumannii, M. tuberculosis*, and *S. aureus,* among others [3-5,53-55]. These studies have been particularly useful for understanding resistance to last resort drugs like vancomycin, colistin, and daptomycin, for which less is known [3,5,52,56,57].

Similar approaches can also be used to identify resistance loci across a library of isolates taken from chronic infections, provided that resistance genes acquire multiple independent mutations across the library. In our studies with *B. dolosa,* we developed a straightforward bacterial genome-wide association study, where we scan each gene for correlations between a phenotypic traits and the presence or absence of mutations that gene. Using this approach, we verified the known role of *gyrA* in ciprofloxacin resistance and identified the gene responsible for O-antigen variation in this strain[6] (Chapter 2). This study did not report any new antibiotic resistance determinants, but the approach of comparing very closely related strains has the potential to do so simply and affordably.

Beyond identifying mutations conferring resistance, tracking the course of antibiotic resistance can reveal new insights into the course of resistance within patients. Snitkin *et al*. additionally tracked the subsequent loss of costly colistin resistance during *A. baumannii* infections of 4 patients. The authors found that in one of these patients, upon removal of drug, *A. baumannii* acquired a mutation that rendered the bacteria less likely to regain colistin resistance[3]. The finding of a mutation that affects future resistance provides hope that we might eventually be able to rationally predict and guide the evolution of bacteria during infection to impede antibiotic resistance.

Bacterial diversification during infection

For many infections, mounting evidence suggests that evolutionary dynamics within the patient are complex. With few exceptions[53], most sequencing studies on sequential isolates from individuals have found evidence that later isolates are not always descendants of earlier isolates. For example, in Smith *et al*.'s 2006 study, targeted sequencing of many *P. aeruginosa* isolated during the 8 years separating the sequenced isolates revealed mutations that were not found in the terminal isolate. Similar genetic evidence of *in vivo* diversification (targeted or whole-genome) has been found in sequential isolates in various infections of the CF lung[6,22,27,58], *E. coli* colonization of the bladder[26], *H. pylori* colonization of the gut[44], and *Mycobacterium tuberculosis* infections[59]. These findings are in agreement with earlier studies of these other infections which have found coexisting variation of phenotypes and particular genetic loci in single patients[29,60-66].

Genomic insights into within-patient diversification

More recently, studies have sought to characterize and date the origin of this co-existing intrastrain diversity using pooled sequencing and independent sequencing of many multiple single colonies cultured from the same time-point[29-31,67,68]. Studies by Didelot *et al*. on *H. pylori*[30] and Harris *et al*. on *S. aureus*[37] have combined the extent of observed diversity with molecular clocks for these species to estimate the time of initial infection, aiding in the detection of transmission events. In our second study of *B. dolosa* evolution, sequencing 29 isolates from the same sputum sample and performing pooled sequencing on single samples from 4 other patients, we inferred that this diversity can be as old as the initial infection in some patients, but that population sweeps may occur in some patients. Furthermore, this study showed that the age of co-existing lineages can be inferred from deep sequencing of population samples, without the need to know mutational linkage[29] (Chapter 3).

These recent studies of diversity have revealed a rapid way to identify selective pressures. In our study of *B. dolosa* intrapatient diversity, we find many cases where diverging lineages underwent parallel evolution within the patient, acquiring independent mutations in the same gene. We even find 4 alleles of the same gene coexisting in single sputum sample[29] (Chapter 3). Remarkably, the only 2 other studies that discuss the identity of mutations separating co-existing pathogen lineages also found evidence of parallel evolution within the patient in *P. aeruginosa*[67] and *M. tuberculosis*[69] infections. Within-patient parallel evolution has also been inferred from sequential isolates[21,58], adding further evidence that this is a common phenomenon and powerful tool for identifying genes under selection *in vivo*. As many diverging lineages can coexist at significantly different frequencies[29,63,69], sequencing many isolates or pooled sequencing may be important for detecting this signature of selection.

So far, the drivers of this diversification process are unknown, though several hypotheses have been proposed. Social cheaters, which gain a growth advantage by not producing a shared common good (e.g. quorum sensing mutants, siderophore mutants), may rise in frequency only in the presence of cooperators[60,70]. Similarly, the immune system, phages, or other bacteria present during infection might selectively target the most abundant lineages, causing frequency-dependent selection and impeding dominance of any one lineage in the population[71,72]. Different niches throughout an organ or the body may select for different survival strategies, a notion supported by differential representation of phages and microbial species across different parts of the lung[13,73]. Yet, these findings of frequent within-patient parallelism suggest that co-existing lineages may be ecologically equivalent; it may be that mutations simply cannot sweep the population due to competition between adapted lineages (clonal interference) and the separation of sub-populations across space. The notion that spatial structure may play an important role in preventing sweeps in some infections is supported by recent findings that granulomas in *M. tuberculosis* infections are started by single bacteria[34]. Alternatively, already differentiated lineages may evolve in parallel against common selective forces. Further studies, which examine the order of mutations acquired in co-existing lineages or identify the location of these different strains across the body, will distinguish between these alternative hypotheses.

Implications of within-patient diversity

Whatever the origin, the finding that bacteria diversify as they adapt to the human body has important implications for clinical diagnosis and research. The full diversity of a pathogen population within the patient must be taken into account when performing antibiotic sensitivity

tests, and, depending on the infection, the typical clinical practice of profiling 3-5 colonies may not be enough to asses the complete diversity[66].

Additionally, as alluded to above, it is often incorrect to assume that sequential isolates obtained during longitudinal studies have descended from one another. Incorrect treatment of an earlier isolate as an ancestor, and thus misorientation of the arrow of mutation, can result in inflated mutation rates. Furthermore, an incorrect inference about the direction of mutational events can mask the role of a mutation in adaptation. For example, in our longitudinal study, we found that the ancestor of the *B. dolosa* outbreak had a stop codon that was reverted 10 independent times over the course of the outbreak to form a functioning glycosyltransferase gene. We found that the full-length gycosyltransferase gene restored O-antigen presentation, suggesting an advantage for O-antigen expression *in vivo* and perhaps a tradeoff during transmission[6] (Chapter 2), consistent with the observation that O-antigen can inhibit adhesion to epithelial cells[74]. Without proper phylogenetic inference, this stop codon would have likely been inferred as the derived allele and the presence of this polymorphism in the population would have been interpreted very differently (Chapter 2).

Diversity and transmission networks

Currently, one of the most common uses of bacterial whole genome is to track transmission networks across the globe and outbreaks[37,75-77]. The presence of within-patient diversity poses both challenges and opportunities for understanding transmission networks. On one hand, the presence of within patient diversity means that the practice of using genomic distance between single isolates to infer transmission networks may be seriously flawed[78,79]. Two isolates from the same patient may be separated by more mutations than isolates from different

patients, and which isolate is chosen will drastically change the topology of the network. Still, sufficiently diverse genomic sequences can be used to identify that patients are part of the same outbreak[3,68,77], to rule out potential transmission events[25], and to suggest that reinfection is less likely than relapse[80,81].

On the other hand, within-patient diversity, if accounted for, can aid in the identification of transmission events. As described above, co-existing diversity can provide a lower bound on time since initial infection, provided that patients are initially colonized by a single clone, which can be a very powerful tool when combined with epidemiological data[28,30,68]. Additionally, a phylogenetic tree that includes many isolates from each patient in an outbreak and that is rooted with an unrelated isolate can help identify transmission events; patients infected by others will have their isolates nested inside of another's diversity[6,36,82]. In theory, this and other approaches can identify transmission even if patients are initially infected with multiple clones or are infected multiple times[28,36]. However many problems with this approach remain: insufficient time to accumulate mutations before transmission may make it impossible to infer the directionality of transmission[6]; frequent parallel evolution at the nucleotide level (relative to generation of other mutations) may cause incorrect phylogenetic inferences; and within-patient fixation or insufficient sampling can wipe away the diversity needed for proper inference[79].

Despite these caveats, within-patient diversity can aid in our understanding of how bacteria spread across organs. In our longitudinal study of *B. dolosa*, we found that multiple isolates in the blood stream of patients were closely related to different lung isolates, suggesting multiple transmission events. This suggests that the transmission from lung to blood represented a general decline in host function rather than multiple transmission events[6] (Chapter 2). Similarly, a study of *E. coli* by Reeves *et al.* was unable to reveal a particular mutation

responsible for the transition between asymptomatic colonization and UTI infection, as the same genotype was also present during colonization[36]. In contrast, in a study by Young *et al*. that tracked the progression from *S. aureus* from carriage to disease, the authors found that all blood isolates, including ones take from multiple draws, were identical. Interestingly, this genotype was significantly different from previous samples from the nasal passageways, suggesting an increased mutation rate or the presence of a distinct unsampled population of *S. aureus* elsewhere in the body[7]. Surely, similar approaches will be used to address important outstanding unknowns in bacterial pathogenesis, such as the directionality of the known transmission between the sinuses and lower airways during CF infections[83].

References

1. Jackson, R.W., Johnson, L.J., Clarke, S.R. & Arnold, D.L. Bacterial pathogen evolution: breaking news. *Trends in Genetics* **27**, 32-40 (2011).

2. Davies, J. & Davies, D. Origins and evolution of antibiotic resistance. *Microbiology and Molecular Biology Reviews* **74**, 417-433 (2010).

3. Snitkin, E.S. *et al*. Genomic insights into the fate of colistin resistance and *Acinetobacter baumannii* during patient treatment. *Genome research* (2013).

4. Devasia, R. *et al*. High proportion of fluoroquinolone-resistant *Mycobacterium tuberculosis* isolates with novel gyrase polymorphisms and a gyrA region associated with fluoroquinolone susceptibility. *Journal of clinical microbiology* **50**, 1390-1396 (2012).

5. Rolain, J.-M. *et al*. Real-time sequencing to decipher the molecular mechanism of resistance of a clinical pan-drug-resistant *Acinetobacter baumannii* isolate from Marseille, France. *Antimicrobial agents and chemotherapy* **57**, 592-596 (2013).

6. Lieberman, T.D. *et al*. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nature genetics* **43**, 1275-1280 (2011).

7. Young, B.C. *et al*. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proceedings of the National Academy of Sciences* **109**, 4550-4555 (2012).

8. Connor, R.I., Sheridan, K.E., Ceradini, D., Choe, S. & Landau, N.R. Change in coreceptor use correlates with disease progression in HIV-1–infected individuals. *The Journal of experimental medicine* **185**, 621-628 (1997).

9. Yates, L.R. & Campbell, P.J. Evolution of the cancer genome. *Nature Reviews Genetics* (2012).

10. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer* **12**, 323-334 (2012).

11. Folkesson, A. *et al*. Adaptation of *Pseudomonas aeruginosa* to the cystic fibrosis airway: an evolutionary perspective. *Nature Reviews Microbiology* (2012).

12. Fitzgerald, J.R. Evolution of *Staphylococcus aureus* during human colonization and infection. *Infection, Genetics and Evolution*.

13. Willner, D. *et al*. Spatial distribution of microbial communities in the cystic fibrosis lung. *ISME J* **6**, 471-4 (2012).

14. Klepac-Ceraj, V. *et al*. Relationship between cystic fibrosis respiratory tract bacterial communities and age, genotype, antibiotics and *Pseudomonas aeruginosa*. *Environmental microbiology* **12**, 1293-1303 (2010).

15. LiPuma, J.J. The changing microbial epidemiology in cystic fibrosis. *Clinical microbiology reviews* **23**, 299-323 (2010).

16. Doggett, R.G., Harrison, G.M. & Wallis, E.S. Comparison of some properties of *Pseudomonas aeruginosa* isolated from infections in persons with and without cystic fibrosis. *Journal of bacteriology* **87**, 427-431 (1964).

17. Govan, J.R. & Deretic, V. Microbial pathogenesis in cystic fibrosis: mucoid *Pseudomonas aeruginosa* and B*urkholderia cepacia*. *Microbiological reviews* **60**, 539-574 (1996).

18. Zlosnik, J.E. *et al*. Mucoid and nonmucoid B*urkholderia cepacia* complex bacteria in cystic fibrosis infections. *Am J Respir Crit Care Med* **183**, 67-72 (2011).

19. Hoboth, C. *et al*. Dynamics of adaptive microevolution of hypermutable P*seudomonas aeruginosa* during chronic pulmonary infection in patients with cystic fibrosis. *Journal of Infectious Diseases* **200**, 118-130 (2009).

20. Hauser, A.R., Jain, M., Bar-Meir, M. & McColley, S.A. Clinical significance of microbial infection and adaptation in cystic fibrosis. *Clinical microbiology reviews* **24**, 29-70 (2011).

21. Smith, E.E. *et al*. Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proceedings of the National Academy of Sciences* **103**, 8487-8492 (2006).

22. Cramer, N. *et al*. Microevolution of the major common *Pseudomonas aeruginosa clones* C and PA14 in cystic fibrosis lungs. *Environmental Microbiology* **13**, 1690-1704 (2011).

23. Huse, H.K. *et al*. Parallel evolution in *Pseudomonas aeruginosa* over 39,000 generations in vivo. *MBio* **1**(2010).

24. Andersson, S.G. & Kurland, C.G. Reductive evolution of resident genomes. *Trends in microbiology* **6**, 263-268 (1998).

25. Price, E.P. *et al*. Within-host evolution of *Burkholderia pseudomallei* over a twelve-year chronic carriage infection. *mBio* **4**, e00388-13 (2013).

26. Zdziarski, J. *et al*. Host imprints on bacterial genomes—rapid, divergent evolution in individual patients. *PLoS pathogens* **6**, e1001078 (2010).

27. Yang, L. *et al*. Evolutionary dynamics of bacteria in a human host environment. *Proceedings of the National Academy of Sciences* **108**, 7481-7486 (2011).

28. Golubchik, T. *et al*. Within-Host Evolution of *Staphylococcus aureus* during Asymptomatic Carriage. *PloS one* **8**, e61319 (2013).

29. Lieberman, T.D. *et al*. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. *Nature genetics* **46**, 82-87 (2014).

30. Didelot, X. *et al*. Genomic evolution and transmission of *Helicobacter pylori* in two South African families. *Proceedings of the National Academy of Sciences* **110**, 13880-13885 (2013).

31. Price, J.R. *et al*. Whole-genome Sequencing Shows That Patient-to-patient Transmission Rarely Accounts for Acquisition of *Staphylococcus aureus* on an Intensive Care Unit. *Clinical Infectious Diseases*, cit807 (2013).

32. Marvig, R.L., Johansen, H.K., Molin, S. & Jelsbak, L. Genome analysis of a transmissible lineage of *Pseudomonas aeruginosa* reveals pathoadaptive mutations and distinct evolutionary paths of hypermutators. *PLoS genetics* **9**(2013).

33. Farhat, M.R. *et al*. Genomic analysis identifies targets of convergent positive selection in drug-resistant *Mycobacterium tuberculosis*. *Nature genetics* **45**, 1183-1189 (2013).

34. Ford, C.B. *et al*. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nature genetics* **43**, 482-486 (2011).

35. Barrick, J.E. *et al*. Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461**, 1243-1247 (2009).

36. Reeves, P.R. *et al*. Rates of mutation and host transmission for an *Escherichia coli* clone over 3 years. *PloS one* **6**, e26907 (2011).

37. Harris, S.R. *et al*. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469-74 (2010).

38. Ford, C.B. *et al*. *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nature genetics* (2013).

39. Croucher, N.J. *et al*. Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430-434 (2011).

40. Pepperell, C.S. *et al*. The Role of Selection in Shaping Diversity of Natural *M. tuberculosis* Populations. *PLoS pathogens* **9**, e1003543 (2013).

41. Falush, D. *et al*. Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proceedings of the National Academy of Sciences 98*, 15056-15061 (2001).

42. Hiller, N.L. *et al*. Generation of genic diversity among *Streptococcus pneumoniae* strains via horizontal gene transfer during a chronic polyclonal pediatric infection. *PLoS pathogens* **6**, e1001108 (2010).

43. Morelli, G. *et al*. Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. *PLoS genetics* **6**, e1001036 (2010).

44. Kennemann, L. *et al*. *Helicobacter pylori* genome evolution during human infection. *Proceedings of the National Academy of Sciences* **108**, 5033-5038 (2011).

45. Oliver, A. & Mena, A. Bacterial hypermutation in cystic fibrosis, not only for antibiotic resistance. *Clinical Microbiology and Infection* **16**, 798-808 (2010).

46. Jolivet-Gougeon, A. *et al*. Bacterial hypermutation: clinical implications. *Journal of medical microbiology* **60**, 563-573 (2011).

47. Oliver, A., Cantón, R., Campo, P., Baquero, F. & Blázquez, J. High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science* **288**, 1251-1253 (2000).

48. Román, F., Cantón, R., Pérez-Vázquez, M., Baquero, F. & Campos, J. Dynamics of long-term colonization of respiratory tract by *Haemophilus influenzae* in cystic fibrosis patients shows a marked increase in hypermutable strains. *Journal of clinical microbiology* **42**, 1450-1459 (2004).

49. Prunier, A.-L. *et al*. High rate of macrolide resistance in *Staphylococcus aureus* strains from patients with cystic fibrosis reveals high proportions of hypermutable strains. *Journal of Infectious Diseases* **187**, 1709-1716 (2003).

50. LeClerc, J.E., Li, B., Payne, W.L. & Cebula, T.A. High mutation frequencies among *Escherichia coli* and *Salmonella pathogens*. *Science* **274**, 1208-1211 (1996).

51. Fricke, W.F. & Rasko, D.A. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nature Reviews Genetics* (2013).

52. Mwangi, M.M. *et al*. Tracking the in vivo evolution of multidrug resistance in *Staphylococcus aureus* by whole-genome sequencing. *Proceedings of the National Academy of Sciences* **104**, 9451-9456 (2007).

53. Saunders, N.J. *et al*. Deep resequencing of serial sputum isolates of *Mycobacterium tuberculosis* during therapeutic failure due to poor compliance reveals stepwise mutation of key resistance genes on an otherwise stable genetic background. *Journal of Infection* **62**, 212-217 (2011).

54. Howden, B.P. *et al*. Evolution of multidrug resistance during *Staphylococcus aureus* infection involves mutation of the essential two component regulator WalKR. *PLoS pathogens* **7**, e1002359 (2011).

55. Howden, B.P. *et al*. Genomic analysis reveals a point mutation in the two-component sensor gene graS that leads to intermediate vancomycin resistance in clinical *Staphylococcus aureus*. *Antimicrobial agents and chemotherapy* **52**, 3755-3762 (2008).

56. Levert, M. *et al*. Molecular and evolutionary bases of within-patient genotypic and phenotypic diversity in *Escherichia coli* extraintestinal infections. *PLoS pathogens* **6**, e1001125 (2010).

57. Tran, T.T. *et al*. Whole-Genome Analysis of a Daptomycin-Susceptible *Enterococcus faecium* Strain and Its Daptomycin-Resistant Variant Arising during Therapy. *Antimicrobial agents and chemotherapy* **57**, 261-268 (2013).

58. McAdam, P.R., Holmes, A., Templeton, K.E. & Fitzgerald, J.R. Adaptive evolution of *Staphylococcus aureus* during chronic endobronchial infection of a cystic fibrosis patient. *PloS one* **6**, e24301 (2011).

59. Merker, M. *et al*. Whole Genome Sequencing Reveals Complex Evolution Patterns of Multidrug-Resistant *Mycobacterium tuberculosis* Beijing Strains in Patients. *PloS one* **8**, e82551 (2013).

60. Wilder, C.N., Allada, G. & Schuster, M. Instantaneous within-patient diversity of *Pseudomonas aeruginosa* quorum-sensing populations from cystic fibrosis lung infections. *Infection and immunity* **77**, 5631-5639 (2009).

61. Israel, D.A. *et al*. *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. *Proceedings of the National Academy of Sciences* **98**, 14625-14630 (2001).

62. Ashish, A. *et al*. Extensive diversification is a common feature of *Pseudomonas aeruginosa* populations during respiratory infections in cystic fibrosis. *Journal of Cystic Fibrosis* (2013).

63. Workentine, M.L. *et al*. Phenotypic Heterogeneity of *Pseudomonas aeruginosa* Populations in a Cystic Fibrosis Patient. *PloS one* **8**, e60225 (2013).

64. Mowat, E. *et al*. *Pseudomonas aeruginosa p*opulation diversity and turnover in cystic fibrosis chronic infections. *American journal of respiratory and critical care medicine* **183**, 1674-1679 (2011).

65. Champion, M.D., Gray, V., Eberhard, C. & Kumar, S. The Evolutionary History of Amino Acid Variations Mediating Increased Resistance of *S. aureus* Identifies Reversion Mutations in Metabolic Regulators. *PloS one* **8**, e56466 (2013).

66. Foweraker, J., Laughton, C., Brown, D. & Bilton, D. Phenotypic variability of *Pseudomonas aeruginosa* in sputa from patients with acute infective exacerbation of cystic fibrosis and its impact on the validity of antimicrobial susceptibility testing. *Journal of Antimicrobial Chemotherapy* **55**, 921-927 (2005).

67. Chung, J.C. *et al*. Genomic Variation among Contemporary *Pseudomonas aeruginosa* Isolates from Chronically Infected Cystic Fibrosis Patients. *Journal of bacteriology* **194**, 4857-4866 (2012).

68. Harris, S.R. *et al*. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *The Lancet infectious diseases* (2012).

69. Sun, G. *et al*. Dynamic population changes in *Mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients. *Journal of Infectious Diseases* **206**, 1724-1733 (2012).

70. Traverse, C.C., Mayo-Smith, L.M., Poltak, S.R. & Cooper, V.S. Tangled bank of experimentally evolved *Burkholderia* biofilms reflects selection during chronic infections. *Proceedings of the National Academy of Sciences* **110**, E250-E259 (2013).

71. Duan, K., Dammel, C., Stein, J., Rabin, H. & Surette, M.G. Modulation of *Pseudomonas aeruginosa* gene expression by host microflora through interspecies communication. *Molecular microbiology* **50**, 1477-1491 (2003).

72. Meyer, J.R. & Kassen, R. The effects of competition and predation on diversification in a model adaptive radiation. *Nature* **446**, 432-435 (2007).

73. Willner, D. *et al*. Case studies of the spatial heterogeneity of DNA viruses in the cystic fibrosis lung. *American Journal of Respiratory Cell and Molecular Biology* **46**, 127-131 (2012).

74. Saldías, M.S., Ortega, X. & Valvano, M.A. *Burkholderia cenocepacia* O antigen lipopolysaccharide prevents phagocytosis by macrophages and adhesion to epithelial cells. *Journal of medical microbiology* **58**, 1542-1548 (2009).

75. Gardy, J.L. *et al*. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *New England Journal of Medicine* **364**, 730-739 (2011).

76. Comas, I. *et al*. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. *Nature genetics* **45**, 1176-1182 (2013).

77. Snitkin, E.S. *et al*. Tracking a hospital outbreak of carbapenem-resistant *Klebsiella pneumoniae* with whole-genome sequencing. *Science translational medicine* **4**, 148ra116-148ra116 (2012).

78. Ypma, R.J., van Ballegooijen, W.M. & Wallinga, J. Relating Phylogenetic Trees to Transmission Trees of Infectious Disease Outbreaks. *Genetics* **195**, 1055-1062 (2013).

79. Didelot, X., Gardy, J. & Colijn, C. Bayesian inference of infectious disease transmission from whole genome sequence data. *bioRxiv* (2013).

80. Bryant, J.M. *et al*. Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. *The Lancet Respiratory Medicine* **1**, 786-792 (2013).

81. Okoro, C.K. *et al*. High-resolution single nucleotide polymorphism analysis distinguishes recrudescence and reinfection in recurrent invasive nontyphoidal *Salmonella typhimurium* disease. *Clinical infectious diseases* **54**, 955-963 (2012).

82. Bryant, J.M. *et al*. Whole-genome sequencing to identify transmission of *Mycobacterium abscessus* between patients with cystic fibrosis: a retrospective cohort study. *The Lancet* (2013).

83. Hansen, S.K. *et al*. Evolution and diversification of *Pseudomonas aeruginosa* in the paranasal sinuses of cystic fibrosis children have implications for chronic lung infection. *The ISME journal* **6**, 31-45 (2011).

Chapter 2:

Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes[1,2]

Bacterial pathogens evolve during the infection of their human hosts[1-8], but separating adaptive and neutral mutations remains challenging[9-11]. Here, we identify bacterial genes under adaptive evolution by tracking recurrent patterns of mutations in the same pathogenic strain during the infection of multiple patients. We conducted a retrospective study of a *Burkholderia dolosa* outbreak among people with cystic fibrosis, sequencing the genomes of 112 isolates collected from 14 individuals over 16 years. We find that 17 bacterial genes acquired non-synonymous mutations in multiple individuals, which indicates parallel adaptive evolution. Mutations in these genes illuminate the genetic basis of important pathogenic phenotypes, including antibiotic resistance and bacterial membrane composition, and implicate oxygen-dependent gene regulation as paramount in lung infections. Several genes have not been previously implicated in pathogenesis, suggesting new therapeutic targets. The identification of parallel molecular evolution suggests key selection forces acting on pathogens within humans and can help predict and prepare for their future evolutionary course.

---

[1] Jean-Baptiste Michel and I contributed equally to this work.
[2] This collaborative work was published in the December 2011 issue of *Nature Genetics* (DOI: 10.1038/ng.997). The authors are Tami D. Lieberman, Jean-Baptiste Michel, Mythili Aingaran, Gail Potter-Bynoe, Damien Roux, Michael R. Davis, Jr., David Skurnik, Nicholas Leiby, John J. LiPuma, Joanna B. Goldberg, Alexander J. McAdam, Gregory P. Priebe, and Roy Kishony.

Introduction

During acute and chronic infections, bacterial pathogens can accumulate mutations that allow them to better adapt to their human host[1,2], evade the immune response[12,13], and become more resistant to antibiotic therapy[3,4]. The spectrum of beneficial mutations that arise during the course of a bacterial infection is likely to reflect genetic pathways critical to bacterial pathogenesis *in vivo*, and therefore may inspire new therapeutic directions. Recent advances in high-throughput sequencing make it possible to follow the genome evolution of bacterial pathogens[7-9,14-17], but it is still difficult to tease apart the adaptive driver mutations from the neutral passenger mutations that have been fixed by chance[9-11]. In the laboratory, this is addressed by following several populations grown in parallel cultures under identical conditions; the adaptive role of mutations is indicated by their recurrence in replicate experiments[17-19]. In natural and clinical environments, such studies are more difficult and have not yet been systematically performed on a genome-wide scale. As a result, global patterns of adaptive evolution that underlie bacterial pathogenesis in humans are not well characterized.

Here, we systematically identify recurrent patterns of evolution implicated in pathogenesis by comparing the genetic adaptation of a single bacterial strain in multiple human subjects during the spread of an epidemic. The airways of people with cystic fibrosis (CF, a lethal genetic disorder) are particularly prone to long-term bacterial infections. Most individuals with CF become colonized by a dominant bacterial strain that persists for many years[20], allowing significant time for genetic adaption[2]. In the 1990s, a small epidemic of *Burkholderia dolosa* – a rare CF pathogen[21,22] that can be transmitted from person to person[23] – broke out among individuals with CF in Boston[24,25]. A total of 39 individuals were infected with *B. dolosa* (Figure

2.1a); and all were followed in a Boston hospital, where bacteria isolated during normal care were routinely frozen.

Results

We conducted a retrospective study of 112 *B. dolosa* isolates from 14 individuals with Cystic Fibrosis from this epidemic outbreak – including the first infected subject in the Boston area (patient zero) – over the course of 16 years (Figure 2.1b and Table S1.1). During this period, five of these individuals received a lung transplant, and eight died. Most of the 112 isolates were recovered from the subjects' airways; a few were recovered from the blood of subjects with bacteremia. This collection covers the epidemic with high temporal resolution and enables us to study the parallel evolution of the same strain in multiple individuals (Figure S1.1).

We sequenced the whole-genome of these 112 *B. dolosa* isolates on an Illumina GAIIx sequencer (75bp single-end reads, average read depth 37x; Figure S1.2) and aligned the reads onto a *B. dolosa* reference genome[26]. We focused our analysis on SNPs; although structural variants and mobile elements may also be important, they are beyond the scope of this study. Our analysis identified 492 polymorphic loci. These mutations accumulated at a steady rate of ~2 SNP/year (r=0.79) (Fig. 1c), with no discernible difference between subjects (Figure S1.3). This rate of mutation accumulation in the presence of selection within the human body is consistent with bacterial mutation fixation rates reported in long-term human infections[2,27]. The steady accumulation of mutations generated enough genetic diversity to resolve evolutionary relationships between isolates, which were investigated through the creation of a maximum-likelihood phylogenetic tree (Figure 2.2a).

Figure 2.1. Whole-genome sequencing of 112 *Burkholderia dolosa* isolates recovered from 14 epidemic patients indicates steady accumulation of mutations over years. (**a**) An epidemic of *B. dolosa* spread to 39 people with cystic fibrosis over decades (circles). Time of first attested infection for each patient is indicated in years since the collection of an isolate from patient zero (labeled 'A'). We studied retrospectively a cohort of 14 patients from this epidemic (gray circles, and labels). (**b)** The genomes of 112 bacterial isolates were sequenced (diamonds; each horizontal line corresponds to a patient). Isolates were recovered over time from the patients' airways (blue), bloodstream (red) or other body compartments (for instance tissue obtained during surgery; orange). (**c)** The number of SNPs between each isolate and the outgroup is plotted as a function of time (years since first isolate). Linear fit is plotted (slope = 2.1 mutations fixed per year).

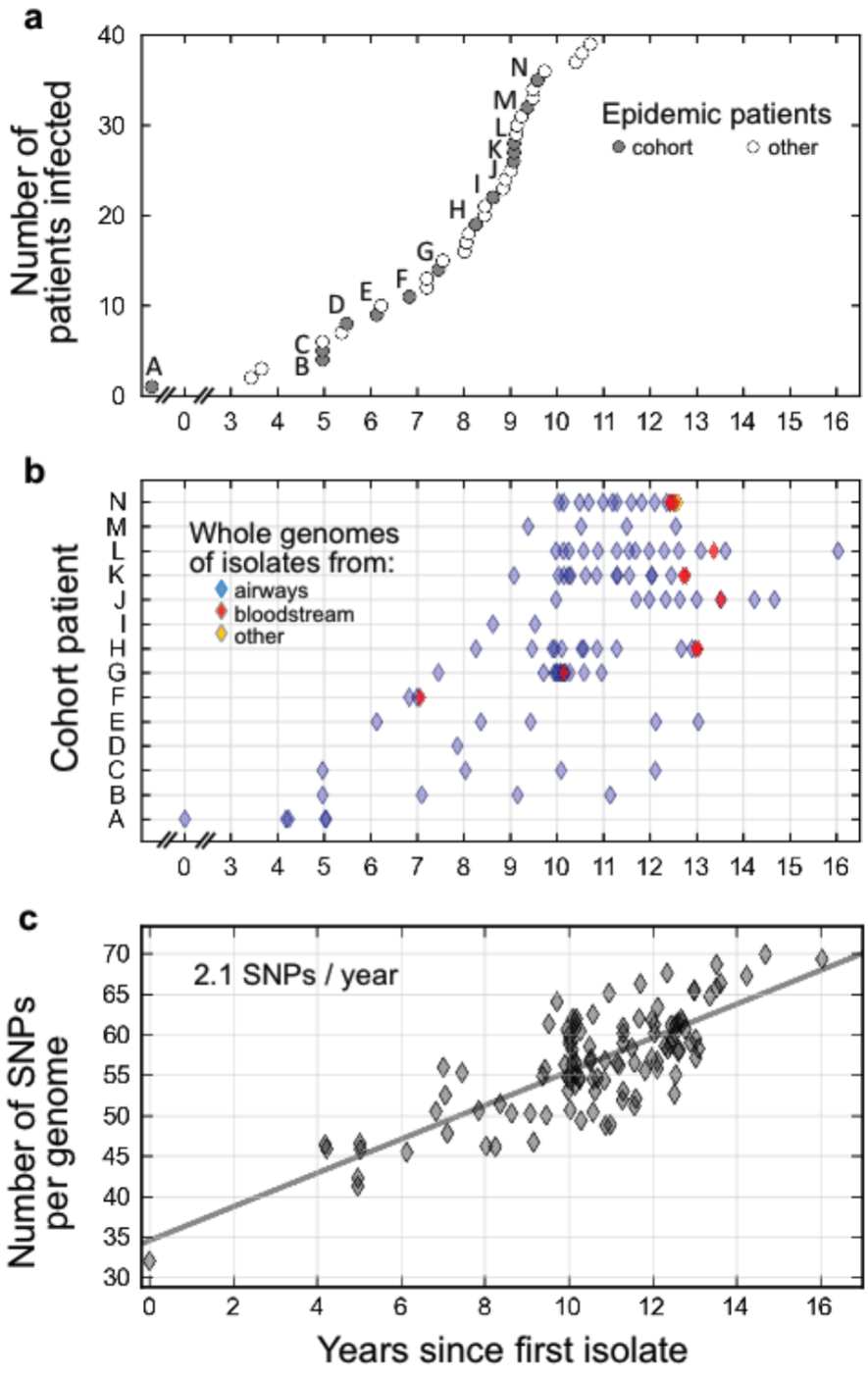

Figure 2.1. Whole-genome sequencing of 112 *Burkholderia dolosa* isolates recovered from 14 epidemic patients indicates steady accumulation of mutations over years (Continued).

Figure 2.2. Bacterial phylogeny reveals a likely network of transmission between patients, and between organs. (a) The maximum-likelihood phylogenetic tree is displayed (SNP scale shown). The 112 isolates are indicated by thin dashed lines colored according to patient, and labeled according to patient and time (e.g., 'C-14-5' was recovered from patient C, fourteen years and five months after the first isolate). Blood isolates are indicated by a dollar sign, and isolates with the same patient and date are distinguished by letters. The last common ancestor (LCA) of isolates from the same patient is represented as a circle of the appropriate color and label. Colored backgrounds indicate patient-specific genetic fingerprints. Patients B, C, E, and F share the same LCA (white background). (**b**) Phylogeny between the inferred LCAs suggests a likely network of infection between patients (arrows). Dashed arrows indicate less certainty (fewer than 3 isolates). **(**c**)**, Phylogeny between blood and lung isolates recovered from the same patient evidences the transmission of multiple clones to the bloodstream during bacteremia (multiple arrows, patients N and K).

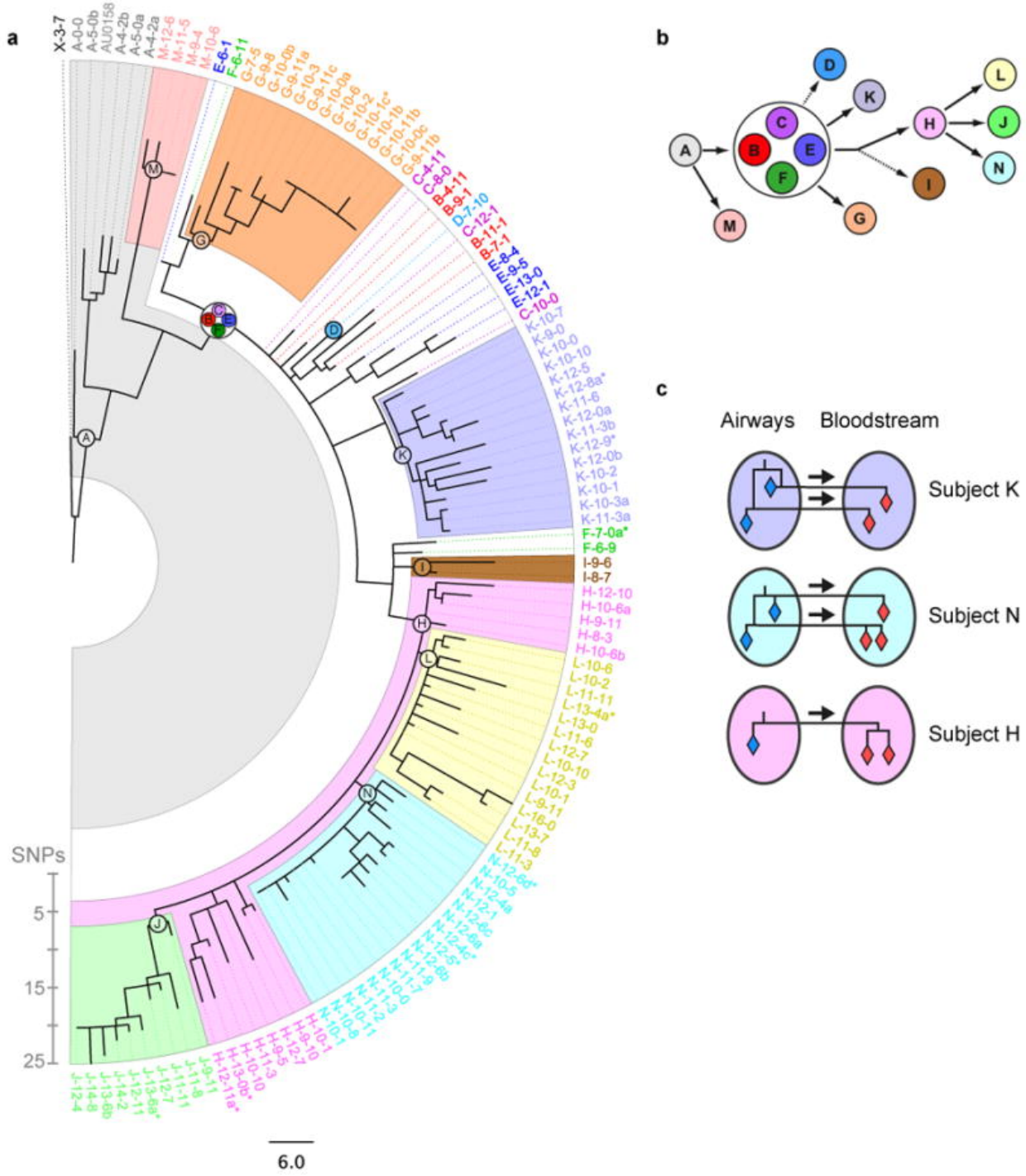

Figure 2.2. Bacterial phylogeny reveals a likely network of transmission between patients, and between organs (Continued)

At the epidemic level, the phylogeny suggests a network of transmission between subjects. Isolates from the same subject tend to form genetically related clusters in the phylogeny (Figure 2.2a, Figure S1.3). These clusters define subject-specific genetic fingerprints, from which transmission history can be inferred. We constructed the last common ancestor (LCA) for each subject's set of isolates, which bears the subject-specific fingerprints. The phylogenetic relationships between these inferred strains indicated the likely network of transmissions among the 14 subjects (Figure 2.2b). Because these data account for only 14 of the 39 subjects in this epidemic, we cannot determine whether transmission occurred directly from one subject to another, or indirectly through a subject not in our study, via a healthcare worker, or through a medical device. Nevertheless, this analysis shows that this specific epidemic was transmitted through several people during its spread and demonstrates the strength of the approach in identifying the infection network of an epidemic.

At the level of individual subjects, the phylogenetic analysis evidenced the transfer of multiple *B. dolosa* clones to the subject's bloodstream during bacteremia. We examined isolates from the three subjects for whom we obtained more than one blood isolate (subjects H, K, and N). In two of these individuals, we found pairs of blood isolates that evolved from distinct lung isolates (Figure 2.2c), which is inconsistent with the transmission of a single clone from lungs to blood (Figure S1.4a). This evidence for multiclonal transmission is consistent either with a punctuated transmission of multiple clones from the genetically diverse lung[28-31] (Figure S1.4b), or with multiple transmissions occurring over time. These different possibilities would lead to recommendations for distinct therapeutic actions: whereas a lung transplant might be effective in preventing the continuous leak of bacteria through lesions of the lung, it would not block the

proliferation of bacteria already within the bloodstream. This analysis thus brings into focus unresolved questions about the mechanistic basis of bacteremia.

Finally, we investigated evolution at the gene level. We looked for genetic correlates of known pathogenic phenotypes. We first assayed resistance to ciprofloxacin, a fluoroquinolone frequently prescribed to CF subjects (Figure 2.3a). Resistance among the 112 isolates varied over two orders of magnitudes (Figure S1.5a). We scored each gene for correlation between the presence of mutations and drug resistance (Figure 2.3a, inset). This genome-wide association study implicated a single gene in the phenotype, BDAG_02180, homolog to *Escherichia coli gyrA*. All the genotypes associated with resistance had nonsynonymous mutations in T83 or D87, known for their role in fluoroquinolone resistance[4,32,33]. Mutations in these residues occurred in six subjects. In each case, phylogenetic analysis indicated that mutations were independently acquired within the subject, after initial infection (Figure S1.5b). In some cases, we even found in the same subject mutations in both residues, each carried by a different isolate. These findings support the presence of a strong selective pressure from fluoroquinolones but suggest that there are only few genetic paths to resistance *in vivo*.

We then focused on a second pathogenic phenotype, the presentation of O-antigen repeats in the lipopolysaccharide (LPS) of the bacterium's outer membrane, known for its importance to virulence in related species[34-36]. We found that some of our isolates present the O-antigen while others do not (Figure 2.3b). A single nucleotide in the glycosyltransferase gene BDAG_02317 correlated exactly with the presentation of O-antigen repeats (Figure 2.3b, inset). The ancestral genotype at this locus, a stop codon, corresponds to the absence of O-antigen repeats; two different mutations at the same amino acid position – each restoring a full-length protein – are associated with presence of the repeats. We confirmed this association

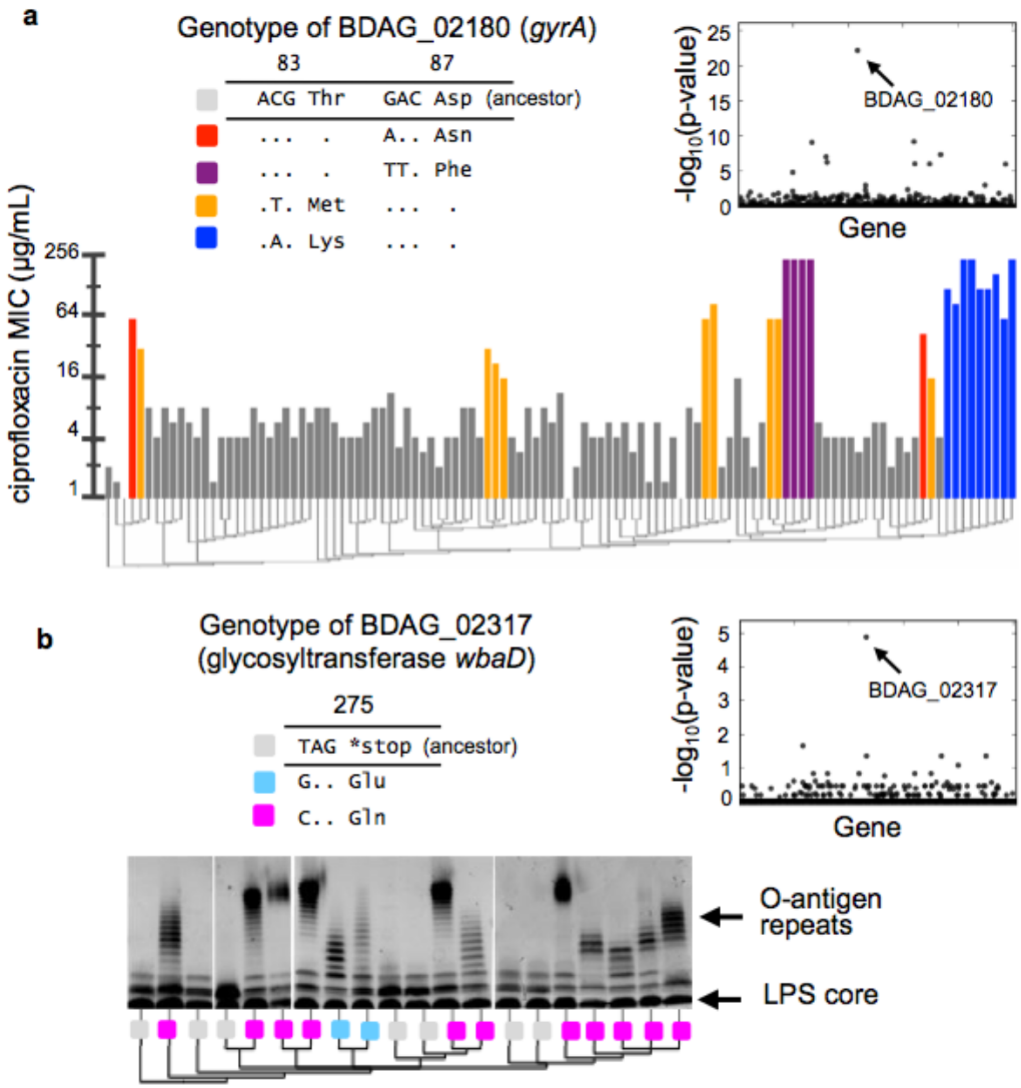

Figure 2.3. Pathogenic phenotypes are tied to point mutations in key genes. (**a**), Minimal inhibitory concentration (MIC) of ciprofloxacin for each isolate (vertical bars) is correlated with genotypic changes in aminoacids 83 and 87 of BDAG_02180, a homolog to *gyrA* (bar colors, legend). Phylogeny is indicated below as a dendrogram. **Inset**, p-values for correlation between the presence of mutations in each gene and resistance levels (Kendall's tau). (**b**), Silver-stained gels showing the presence of O-antigen repeats (banded pattern) in the LPS of eight isolates. Genotypes of BDAG_02317 (a glycosyltransferase) are shown in the legend. The presentation of O-antigen repeats corresponds to a recurrent gain-of-function mutation.

experimentally; we found that complementation with the full-length glycosyltransferase gene could restore O-antigen presentation (Figure S1.6a, Supplementary Information 1). Harnessing the phylogenetic information, we determined that the last common ancestors of strains from each subject presented the truncated genotype. Thus, the gain-of-function mutations occurred in nine subjects independently (Figure S1.6b), highlighting the strength of the selective pressure for O-antigen presentation during the infection. These results identify a previously uncharacterized genetic mechanism for O-antigen switching and hint at a tradeoff during person-to-person transmission.

We recognize that the human body challenges bacteria with many selective pressures beyond those discussed above. We therefore developed a systematic approach for identifying genes under positive selection without prior knowledge of the phenotypes being selected. At the genome level, we found no evidence for selection in coding regions (dN/dS~1) and no significant intragenic bias (Supplementary Information 1). However, we reasoned that genes under selection would be mutated independently in different subjects[17-19]. We leveraged the phylogeny to calculate the number of mutations each gene received, distinguishing genes mutated multiple times from those mutated once but appearing in several subjects through the expansion of the lineage that carried them. We counted 561 independent mutational events in 304 genes. Assuming neutral evolution, we would expect that these mutations would distribute randomly among the 5,014 *B. dolosa* genes, and that genes would rarely acquire more than a single mutation. Instead, we observed that many more genes than expected contained multiple mutations (Figure 2.4a, inset). Seventeen genes were found to have three or more different mutations (neutral expectation: ~1, Methods), and four genes had over ten different mutations (expected: 0 genes).

To determine whether genes that acquired multiple mutations were under positive selection, or were merely sites of mutational bias, we calculated the canonical measure for selection, dN/dS (Figure 2.4a). The large subset of 247 genes which contained only one mutation exhibited a weak but significant signal for purifying (i.e., negative) selection (dN/dS=0.63, $p<10^{-3}$, Methods). The set of genes with two mutations did not show evidence of selection (dN/dS=1.4, CI:0.7-3.1); this set may include a combination of genes that are under some selection and genes that fixed two mutations by chance (22 expected under neutral drift; 28 observed). By contrast, the 17 genes that acquired three or more mutations received 18 times as many non-synonymous mutations as expected by neutral drift, and are under strong positive selection (dN/dS=18, CI:4.9-152.7). This suggests that these seventeen genes are not neutral mutational hotspots; they are undergoing adaptive evolution under the pressure of natural selection.

The 17 genes under positive selection (Figure 2.4b, Table S1.2), which are mostly conserved across the *Burkholderia* genus (Figure S1.7), indicate genetic pathways that may be involved in pathogenesis. The presence of the two genes previously identified in connection with antibiotic resistance and O-antigen presentation (*gyrA*: 11 mutations; glycosyltransferase *wbaD*: 10 mutations) provides a further connection of these genes to pathogenic phenotypes under selection. Eleven of the 17 genes belong to functional categories related to pathogenicity: membrane synthesis (4 genes, including 2 in LPS biosynthesis), secretion (2), and antibiotic resistance (5). The presence of a second glycosyltransferase in the O-antigen cluster (6 mutations) stresses the importance of this pathway to the disease. Notably, the remaining 6 genes had not previously been implicated in pathogenesis of lung infections. Three of these – a glucoamylase, a methyltransferase, and a sigma factor – have no well-annotated close homologs and their roles in pathogenicity are thus unclear. Another gene trio (homologs of *fnr, fixL*, and

Figure 2.4. Parallel evolution identifies a set of genes under strong selection during pathogenesis. (a), **Inset**, The number of genes that acquired at least *m* mutations across the epidemic is plotted as a function of *m* (gray bars). This distribution contrasts sharply with the distribution expected for neutral evolution (black line). Under neutral evolution, the expectation is that only one gene would receive three or more mutations (m>3); instead, we observed 17 such genes. **Main**, The canonical signal for selection (dN/dS) is calculated for these 17 genes (109 mutations in these 17 genes), the 28 genes with m=2, and the 247 genes with m=1 (**Supplementary Fig. 3** presents the contribution of these mutations to the molecular clock). Values of dN/dS greater than 1 indicate positive selection (blue), values smaller than 1 indicate purifying selection (red). Error bars indicate 95% CIs. Calculated over all genes without regard to *m*, this analysis would not show a signal for selection. (**b**), Each one of the 17 genes (rows) under positive selection contained an acquired mutated in several patients, signified by squares (color intensity indicate the number of mutations observed within this patient). The total number of mutations observed within that gene, *m*, is indicated left. Genes are grouped by biological function, and labeled with the annotations of close homologs, and, when available, close homolog names.

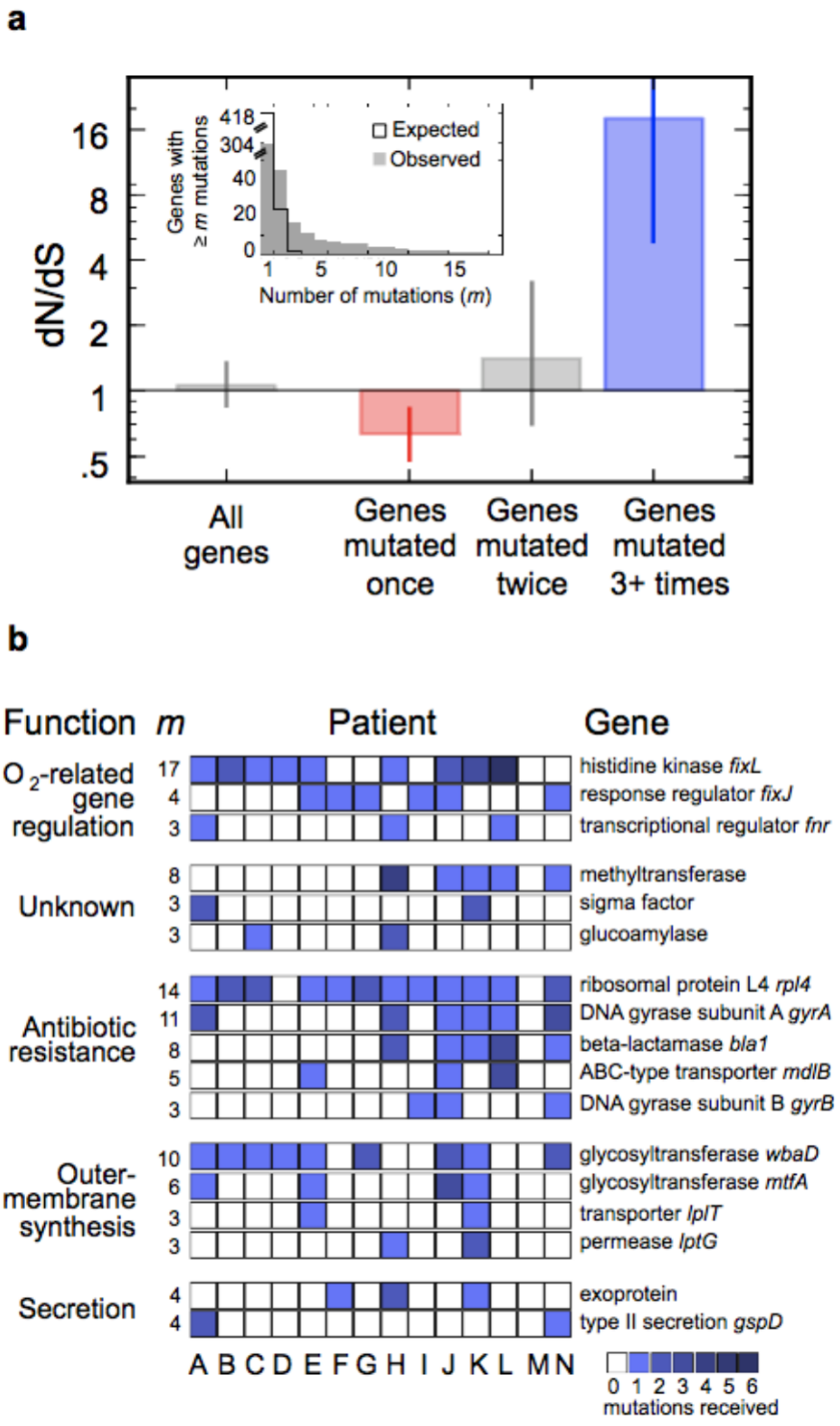

Figure 2.4. Parallel evolution identifies a set of genes under strong selection during pathogenesis (Continued).

fixJ), including the gene most mutated gene (BDAG_01161, a homolog of *fixL,* that had 17 nonsynonymous mutations), can be linked through homology to oxygen-dependent gene regulation[37]. The large number of mutations in this pathway resonates with reports of lowered oxygen tension in CF mucus[38] and of ties between oxygen sensing and virulence modulation in the gastrointestinal tract[39]. Homologs of these three genes have been implicated in diverse regulatory processes[37,39], but their function and the genes they regulate in *B. dolosa* are currently unknown. The identification of 17 *B. dolosa* genes that underwent selective pressure during infection in subjects with cystic fibrosis highlights key pathways involved in pathogenesis and may suggest new therapeutic targets for this and other lung infections.

Tracking the genomic evolution of bacterial pathogens during the infection of their human host provides a direct method for observing evolutionary mechanisms *in vivo* and identifying genes central to pathogenesis. This study, which harnesses the combination of high-throughput sequencing and parallel evolution in the clinical settings, is a step towards a comprehensive understanding of genetic adaptation during pathogenesis. Systematically identifying selective pressures acting on pathogens within their hosts may help point to new therapeutic directions.

Methods

Study cohort and bacterial isolates,

An epidemic of Burkholderia dolosa affected 39 patients in the Boston area over twenty years[25,26]. Our cohort includes patient zero, the seven patients for whom bacterial isolates recovered from the bloodstream were available, and six patients chosen at random (Supplementary Information 1). Samples from these patients were collected during normal care (Table S1.1) and frozen. Frozen clinical stocks were streaked on solid media; a single colony from each plate was chosen at random and frozen in 15% glycerol to create a working library. Time of isolation is reported relative to the collection of an isolate from patient zero (isolate A-0-0). The use of discarded samples for this study was approved by the institutional review boards at Children's Hospital Boston and Harvard Medical School.

Genome sequencing and SNP calling

DNA was extracted from single colonies using standard procedures, and multiplexed genomic libraries were constructed using the Illumina-compatible Nextera™ DNA Sample Prep Kit. Sequencing was performed with 75-bp, single-end reads at a mean read depth of 37x. Reads were aligned using the *B. dolosa* genome AU0158 as a reference, which was isolated from patient zero. We used SAMtools 0.1.12a to manipulate consensus sequences and find SNP between isolates. Details can be found in the Supplementary Information 1, Table S1.1, Figure S1.1, and Figure S1.2.

Rate of evolution

We estimated the number of SNPs accumulated between each isolate and the outgroup, normalizing the SNPs called with confidence to the portion of the genome covered with equally high confidence (Supplementary Information 1). These SNPs were found to accumulate at the constant rate of 2.1 SNPs/year (Pearson r=0.79), corresponding to a mutation fixation rate of $3*10^{-7}$ per base pair per year. Within patients, mutations accumulate at a similar rate (Figure S1.3).

Phylogeny

The single-nucleotide polymorphisms were used to construct the maximum likelihood phylogenetic tree between the 112 isolates (Supplementary Information 1). Our model assumes independent evolution at each site, and vertical inheritance. We used the software implementation Dnaml (Phylip v3.69[39]). Different transition-to-transversion ratios produced remarkably similar phylogenies; we therefore chose a default model where all mutations were equally likely. The LCA for strains from each subject was estimated as the most outward node from which all isolates from that patient descended. Each LCA is an inferred genome that contains all polymorphisms shared among isolates of that patient.

Ciprofloxacin resistance assay

Bacterial isolates were grown at 37°C shaking for 24 hours in microtiter plates containing LB with logarithmically increasing concentrations of ciprofloxacin (cat # 17850, Sigma-Aldrich, USA, Concentrations: 128, 64, 32, 16, 8, 4, 2, 1, 0.5, 0 mg/L; we used 10mM hydrochloric acid for solubilizing a drug stock of 640 mg/L). The minimum inhibitory concentration (MIC) of an

isolate was estimated as the lowest concentration of ciprofloxacin at which growth was < 10% of the maximum growth of that isolate in the absence of drug, as determined by optical density. The reported value for each isolate is the logarithmic average of two replicate experiments performed on different days.

O-antigen repeat assay

Twenty isolates (all taken from the airway, including 5 sets of isolates taken from the same patient, see Supplementary Information 1) from our library were assayed for O-antigen presence. Lipopolysaccharide (LPS) was extracted from each sample as described elsewhere[30]. Extracts were run on SDS-PAGE gels and visualized using Pro-Q Emerald staining. The presence of low molecular weight bands on the gel indicated O-antigen repeats. See Supplementary Information 1, Table S1.3, and Figure S1.6a for details on O-antigen complementation.

Genome-wide association study to detect genetic correlates of assayed pathogenic phenotypes

For each phenotypic assay (MIC to ciprofloxacin; presence/absence of O-antigen repeats), and for each gene, we use our 112 strain library to calculate the correlation between the value of the phenotype and the presence of SNP (with respect to the LCA of the epidemic). Significance was assessed using Kendall's tau in the case of antibiotic resistance, and with Fisher's exact test in the case of the presentation of O-antigen repeat.

Number of mutations

The presence of the same SNP in two distinct isolates can signify one of two events: the mutation occurred once in an ancestral isolate and was passed on to its descendants, or the mutation occurred twice. We resolved this ambiguity by using the phylogenetic tree derived above, inferring the genotypes of all internal nodes using parsimonious assumptions. This method evidences 20 nucleotide positions that were mutated more than once, including 8 positions that mutated 3 or more times. We find 8 positions that mutated to 2 different nucleotides, and 1 position where all 4 nucleotides were observed. Overall, we count 561 mutations. For each gene mutated multiple times, we report the number of mutations that each patient received in Figure 2.4b. For each nucleotide position within the gene, a mutation is counted if there is polymorphism within that patient, or if a mutation first appears in that patient. In this way, we do not count inherited mutations, which arose earlier in the epidemic.

Distribution of mutations per gene expected by random drift

We randomly draw 561 positions in the *B. dolosa* reference genome, and count the distribution of mutations per gene obtained. We repeat this procedure 1000 times and average the results.

Estimating the strength of selective pressures

All but 14 of the 304 genes with polymorphisms correspond to regions of the reference genome with no frameshift mutation: we can assess, for these 290 genes, whether mutations change the translated aminoacid (nonsynonymous, N) or have no such effect (synonymous, S). For any subset of the 290 genes, it is possible to count the observed N and S mutations, and

compare these numbers with those expected under random drift (expected N/S: 2.97; robust to transition-to-transversion ratio). The corresponding dN/dS (ratio of substitution rates at nonsynonymous and synonymous sites) indicates whether selection might be acting on the group of genes under study. Confidence is estimated with Clopper-Pearson binomial confidence intervals (95% confidence); one-tailed p-values are computed by simulation.

Manual annotation of genes

We manually annotated the 17 genes found to be under strong positive selection (Table S1.2). Their translated coding sequences were blasted against the ref_seq database using the BLASTP algorithm on the NCBI website. Well-annotated proteins (3-4 letter gene name and a link to NCBI gene) with high homology (E value < 10^{-20}) were recorded. If the homology covered < 95% of the protein sequence, or if there was no such well-annotated protein, the protein with highest homology (lowest E value) was recorded. Exceptionally, the gene *Shigella flexneri fnr* gene (not in RefSeq) was included in the annotation list based on 89% homology coverage and the presence of two other oxygen-associated genes (*fixJ* and *fixL*) among the 17 genes under strong positive selection.

Contributions

JBM, AJM and RK conceived the study. JJL, AJM and GPP collected the clinical samples. TDL and NL performed resistance phenotyping. JBG, DR, MRD, DS, and GPP performed LPS phenotyping and complementation. MA, GPB, AJM and GPP conducted chart review and provided medical information. TDL, JBM and RK performed whole-genome sequencing and data analysis. TDL, JBM, JJL, AJM, GPP and RK interpreted the results and wrote the manuscript.

Acknowledgements

We are grateful to M. Caimano, M. Cendron, P. Kokorowski, S. Lory, C. Marx, N. Delany, S. Walker, M. Waldor and R. Ward for insightful discussions and comments, to O. Iartchouk, A. Brown, M. Light and their team at Partners HealthCare Center for Personalized Genetic Medicine for Illumina sequencing, to J. Deane and L. Williams for technical assistance, to S. Vargas for assistance with IRB protocols and to M. Baym, M. Ernebjerg, A. Palmer, E. Toprak, K. Vetsigian, Z. Yao and all of the Kishony lab members for helpful discussions and general support. JBM was supported by the Foundational Questions in Evolutionary Biology Prize Fellowship and the Systems Biology PhD Program (Harvard Medical School). GPP was supported in part by The Mannion Fund for Research of the Center for the Critically Ill Child of Children's Hospital Boston. JJL was supported by the Cystic Fibrosis Foundation. This work was supported in part by US National Institutes of Health grants (GM080177 to the Systems Biology Department, Harvard Medical School and GM081617 to RK), by a grant from the New England Regional Center of Excellence for Biodefense and Emerging Infectious Diseases (NERCE; AI057159 to RK) and by a Harvard Catalyst grant (to RK, AJM. and M. Cendron).

References

1. Suerbaum, S. & Josenhans, C. *Helicobacter pylori* evolution and phenotypic diversification in a changing host. *Nat Rev Microbiol* **5**, 441-452 (2007).

2. Smith, E. E. *et al*. Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proc Natl Acad Sci U S A* **103**, 8487-8492 (2006).

3. Musher, D. M. *et al*. Emergence of macrolide resistance during treatment of pneumococcal pneumonia. *N Engl J Med* **346**, 630-631 (2002).

4. Wong, A. & Kassen R. Parallel evolution and local differentiation in quinolone resistance in *Pseudomonas aeruginosa*. *Microbiology* **157**, 937-944 (2011).

5. Zdziarski, J. *et al*. Host imprints on bacterial genomes—rapid divergent evolution in individual patients. *PLoS Path* **6**, e1001078 (2010).

6. Yang, L. *et al*. Evolutionary dynamics of a bacteria in a human host enviornment. *Proc Natl Acad Sci U S A* **108**, 7481-7486 (2011).

7. Kennemann, L. *et al*. *Helicobacter pylori* genome evolution during human infection. *Proc Natl Acad Sci U S A* **108**, 5033-5038 (2011).

8. Mwangi, M. M. *et al*. Tracking the in vivo evolution of multidrug resistance in *Staphylococcus aureus* by whole-genome sequencing. *Proc Natl Acad Sci U S A* **104**, 9451-9456 (2007).

9. Harris, S. R. *et al*. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469-474 (2010).

10. Goodarzi, H., Hottes, A. K. & Tavazoie, S. Global discovery of adaptive mutations. *Nat Methods* **6**, 581-583 (2009).

11. Pleasance, E. D. *et al*. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191-196 (2010).

12. Moxon, E. R., Rainey, P. B., Nowak, M. A. & Lenski, R. E. Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr Biol* **4**, 24-33 (1994).

13. van der Woude, M. W. & Baumler, A. J. Phase and antigenic variation in bacteria. *Clin Microbiol Rev* **17**, 581-611, table of contents (2004).

14. Croucher, N. J. *et al*. Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430-434 (2011).

15. Holt, K. E. *et al*. High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*. *Nat Genet* **40**, 987-993 (2008).

16. Pallen, M. J. & Wren, B. W. Bacterial pathogenomics. *Nature* **449**, 835-842 (2007).

17. Elena, S. F. & Lenski, R. E. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet* **4**, 457-469 (2003).

18. Woods, R. *et al*. Tests of parallel molecular evolution in long-term experiment with *Escherichia coli*. *Proc Natl Acad Sci U S A* **103**, 9107-9112 (2006).

19. Barrick, J. E. *et al*. Genome evolution and adpatation in a long-term experiment with *Escherichia coli*. *Nature* **461,** 1243-1247 (2009).

20. Lipuma, J. J. The changing microbial epidemiology in cystic fibrosis. *Clin Microbiol Rev* **23**, 299-323 (2010).

21. Vermis, K. *et al*. Proposal to accommodate *Burkholderia cepacia* genomovar VI as *Burkholderia dolosa* sp. nov. *Int J Syst Evol Microbiol* **54**, 689-691 (2004).

22. Lipuma, J. J. Update on the *Burkholderia cepacia* complex. *Curr Opin Pulm Med* **11**, 528-533 (2005).

23. LiPuma, J. J., Dasen, S. E., Nielson, D. W., Stern, R. C. & Stull, T. L. Person-to- person transmission of *Pseudomonas cepacia* between patients with cystic fibrosis. *Lancet* **336**, 1094-1096 (1990).

24. Biddick, R., Spilker, T., Martin, A. & LiPuma, J. J. Evidence of transmission of *Burkholderia cepacia*, *Burkholderia multivorans* and *Burkholderia dolosa* among persons with cystic fibrosis. *FEMS Microbiol Lett* **228**, 57-62 (2003).

25. Kalish, L. A. *et al*. Impact of *Burkholderia dolosa* on lung function and survival in cystic fibrosis. *Am J Respir Crit Care Med* **173**, 421-425 (2006).

26. *Burkholderia dolosa* Sequencing Project. Broad Institute of Harvard and MIT (http://www.broadinstitute.org/).

27. Morelli, G. *et al*. Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. *PLoS Genet* **6**, e1001036 (2010).

28. Sibley, C. D. *et al*. A polymicrobial perspective of pulmonary infections exposes an enigmatic pathogen in cystic fibrosis patients. *Proc Natl Acad Sci U S A* **105**, 15070-15075 (2008).

29. Guss A. M. *et al*. Phylogenetic and metabolic diversity of bacteria associated with cystic fibrosis. *ISME J* **5**, 20 (2011).

30. Mowat, E., *et al*. *Psuedomonas aeruginosa* population diversity and turnover in cystic fibrosis infections. *Am J Respir Crit Care Med* **183**, 1674-1679 (2011).

31. Wilder, C. N., Allada G., & Schuster, M. Instantaneous within-patient diversity of *Psuedomonas aeruginosa* quorum-sensing populations from cystic fibrosis lung infections. *Infect Immun* **77**, 5631-5639 (2009).

32. Weigel, L. M., Steward, C. D. & Tenover, F. C. *gyrA* mutations associated with fluoroquinolone resistance in eight species of *Enterobacteriaceae*. *Antimicrob Agents Chemother* **42**, 2661-2667 (1998).

33. Reyna, F., Huesca, M., Gonzalez, V. & Fuchs, L. Y. *Salmonella typhimurium gyrA* mutations associated with fluoroquinolone resistance. *Antimicrob Agents Chemother* **39**, 1621-1623 (1995).

34. Silhavy, T. J., Kahne, D. & Walker, S. The bacterial cell envelope. *Cold Spring Harb Perspect Biol* **2**, a000414 (2010).

35. Vinion-Dubiel, A. D. & Goldberg, J. B. Lipopolysaccharide of *Burkholderia cepacia* complex. *J Endotoxin Res* **9**, 201-213 (2003).

36. Ortega, X. *et al*. Reconstitution of O-Specific Lipopolysaccharide Expression in *Burkholderia cenocepacia* Strain J2315, Which Is Associated with Transmissible Infections in Patients with Cystic Fibrosis. *J Bact* **187**, 1324-1333 (2005).

37. Crosson, S., McGrath, P. T., Stephens, C., McAdams, H. H. & Shapiro, L. Conserved modular design of an oxygen sensory/signaling network with species- specific output. *Proc Natl Acad Sci U S A* **102**, 8018-8023 (2005).

38. Worlitzsch, D. *et al*. Effects of reduced mucus oxygen concentration in airway *Pseudomonas* infections of cystic fibrosis patients. *J Clin Invest* **109**, 317-325 (2002).

39. Marteyn, B. et al. Modulation of *Shigella* virulence in response to available oxygen in vivo. *Nature*. **465**, 355-358 (2010).

Chapter 3:

Genetic variation of a bacterial pathogen within multiple patients provides a record of past selective pressures[3]

Advances in sequencing have enabled the identification of mutations acquired by bacterial pathogens during infection[1-10]. However, it remains unclear whether adaptive mutations fix in the population or lead to pathogen diversification within the patient[11,12]. Here, we study the genotypic diversity of *Burkholderia dolosa* within people with cystic fibrosis by re-sequencing individual colonies and whole populations from single sputum samples. Extensive intra-sample diversity reveals that mutations rarely fix within a patient's pathogen population—instead, diversifying lineages coexist for many years. When strong selection is acting on a gene, multiple adaptive mutations arise but neither sweeps to fixation, generating lasting allele diversity that provides a recorded signature of past selection. Genes involved in outer-membrane components, iron scavenging and antibiotic resistance all showed this signature of within-patient selection. These results offer a general and rapid approach for identifying selective pressures acting on a pathogen in individual patients based on single clinical samples.

---

[3] This collaborative work was published in the January 2014 issue of *Nature Genetics* (DOI: 10.1038/ng.2848). The authors are Tami D. Lieberman, Kelly B. Flett, Idan Yelin, Thomas R. Martin, Alexander J. McAdam, Gregory P. Priebe, and Roy Kishony.

Introduction

Two opposing models of within-patient bacterial evolution have been proposed*:* a "dominant lineage" model, in which beneficial mutations drive superior lineages to dominate in the population, and a "diverse community" model whereby adaptive lineages rise to intermediate frequency and coexist with other lineages (Figure 3.1)[11-14]. The diversity of within-patient pathogen populations has major implications for drug treatment and resistance[7,15,16], for inferring transmission networks[8,9,17,18], and for understanding evolutionary processes[13,19]. Here, to distinguish between these models and to understand the sources of genetic diversity, we compared the genomes of many bacterial cells of the same strain from the same clinical sample.

We focused on chronic infections with *Burkholderia dolosa,* a rare and deadly opportunistic pathogen that spread among 39 people in with cystic fibrosis (CF) cared for at a single center in Boston starting in the 1990s[20,21]. The airways of these patients were infected with very similar starting strains, and surviving patients have been colonized for many years. A previous retrospective study of single-colony isolates revealed specific *B. dolosa* genes that evolved under strong selective pressures during the outbreak[8]. Now, using sputum samples collected during clinical care, we characterize contemporary intraspecies diversity in 5 individuals from this outbreak who have been infected with *B. dolosa* since the early 2000's.

Results

We used two genomic approaches, colony re-sequencing (Patient 1) and deep population sequencing (Patients 1-5), to identify single nucleotide mutations and their frequencies within single sputum samples. In our colony re-sequencing approach, we isolated dozens of colonies from a clinical sample and analyzed their genomes individually by alignment of reads to a *B.*

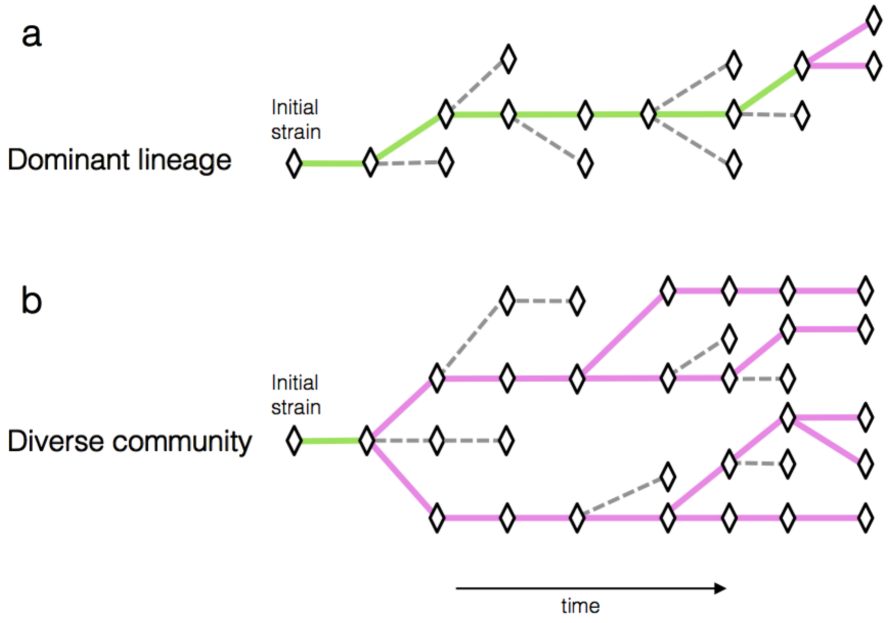

Figure 3.1 Alternative models of within-patient evolution. (a) In the dominant-lineage model of within-host evolution, lineages with beneficial mutations sweep to fixation (green lines), eliminating their less fit ancestors or other temporarily arising genotypes (dashed lines). In this model, most observed mutations will be fixed and polymorphic mutations will be rare, representing only recent mutational events (magenta lines). **(b)** In the diverse community-model, lineages coexist and compete for long stretches of time. In this model, most sampled mutations will be polymorphic.

dolosa reference genome, AU0158, a strain taken from a different patient in this outbreak. Since each colony originates from a single bacterium, this approach is equivalent to comparing different bacterial cells from the initial clinical sample. In the population sequencing approach[22,23], we pooled hundreds of colonies from each clinical sample and sequenced the pool with deep coverage (~450x). We then aligned reads to AU0158 and identified fixed mutations, appearing in all reads, and polymorphisms, appearing in only a fraction of the reads. To remove false positive polymorphic sites caused by systematic sequencing or alignment errors[24,25], we developed a set of thresholds and statistical tests that reject polymorphic sites where the mutated and ancestral reads have significantly different properties[22,23] (see Supplementary Information 2). We calibrated this approach using an isogenic control for which we expect no polymorphisms. For validation, we performed both methods on a single sample from Patient 1, comparing diversity among 29 individual colonies to the population sequencing approach (Figure 3.2). The population sequencing approach reliably detects polymorphisms where the minor allele frequency is larger than 3%, while decreasing the cost and labor required per sample.

We found that most mutations that arise during the course of infection do not fix, but remain polymorphic within the patient. The colony re-sequencing approach performed for Patient 1 revealed 188 mutations occurring in some, but not all, isolates and only 10 mutations shared among all isolates. This dominance of polymorphisms, also seen in the population sequencing from the same sample, strongly supports the diverse community model (Figure 3.3a-b). Similarly, for the four other patients, population sequencing on single samples identified a preponderance of polymorphisms compared to fixed mutations (≥73% of mutations, Figure 3.3b). We found these excesses despite the bias to overestimate mutations fixed during infection;

Figure 3.2: Two methods for studying genomic intraspecies diversity. (**a**) To study within-patient evolution, we cultured sputum samples from patients with cystic fibrosis on selective media. In the colony re-sequencing approach (solid arrows, performed for one patient), we isolated multiple individual colonies from the same single sample, independently called variants for each isolate via alignment of reads, and compared variants among the isolates. In the deep population sequencing approach (dashed arrow, performed for five patients), we pool hundreds of colonies from the same plate and analyze the pool's genomic DNA. We identified positions on the genome where some reads, originating from different colonies on the plate, disagree with an inferred ancestral genome (Methods). (**b**) Allele frequency estimates in the population sequencing (y-axis) versus the colony re-sequencing (x-axis) from the same sputum sample (P1) for each mutated position. Mutations are classified as either fixed (green circles) or polymorphic (magenta circles). Some mutations found in the colony-based approach are sub-threshold in frequency or confidence in the pool-based approach (open squares). Slight jitter is added in the X and Y locations for each point to improve visibility (up to 2% change). As an example, the insets at top and at right display a summary of the raw data at the indicated genomic position. The population sequencing (right) at this position shows 70% aligned reads supporting a T (orange) and 30% supporting a G (black), consistent with the corresponding number of colonies in the individual isolates (22, T; 7, G). Reads from each isolate (top) are mostly of identical calls (all T, or all G). Green indicates a single read in one isolate supporting an A, likely a sequencing error. For further comparison of the two methods, see Figure S2.7.

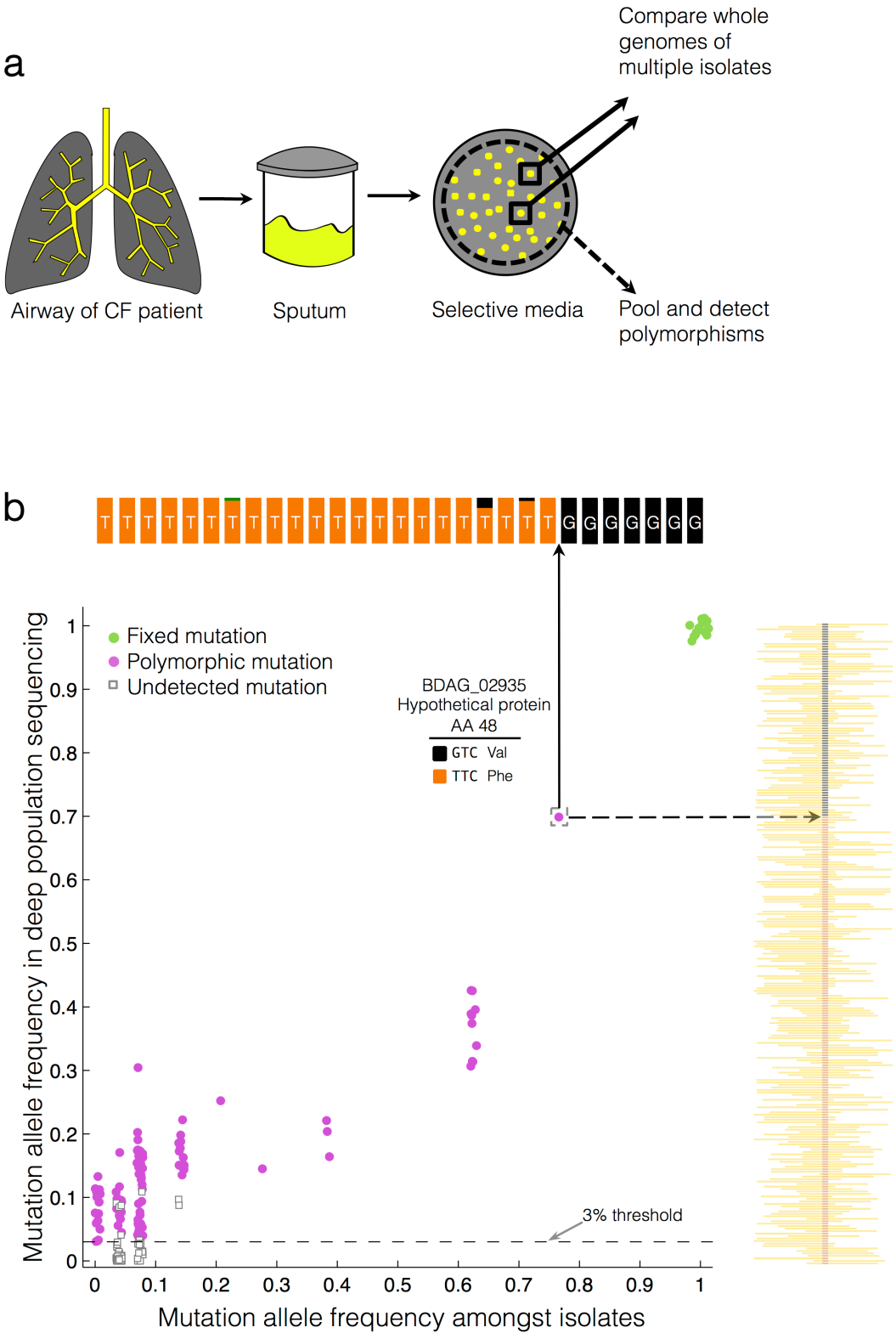

Figure 3.2: Two methods for studying genomic intraspecies diversity (Continued).

Figure 3.3: Within-patient evolution leads to diversification, not substitution. Mutations found in *B. dolosa* within-patient populations relative to the outgroup are classified as fixed (green), or polymorphic (magenta). An excess of polymorphic versus fixed mutations supports the diverse-community model over the dominant-lineage model. (**a**) A maximum-parsimony phylogeny of 29 isolates from the same sputum sample (P1) shows the coexistence of diverse sub-lineages separated by many single nucleotide mutations accumulating since the last common ancestor (LCA) of this patient. Each isolate is represented by a dotted line. (**b**) The diverse-community model is also supported by the distribution of allele frequencies from the population sequencing in 5 patients' samples.

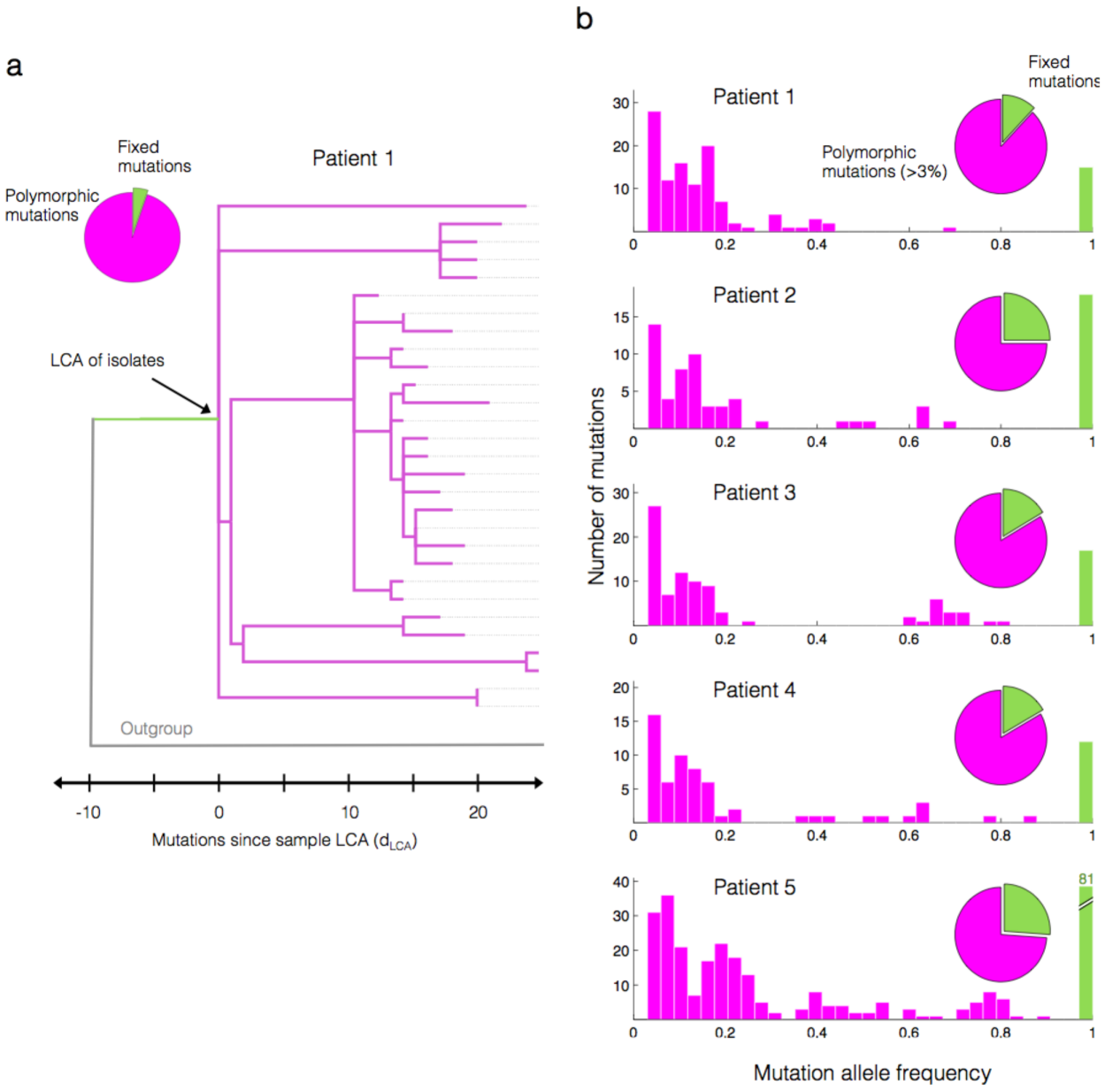

Figure 3.3: Within-patient evolution leads to diversification, not substitution (Continued).

some fixed mutations in a sputum sample might be polymorphic within the patient's airways or have fixed prior to patient colonization (Figure S2.1).

The observed genomic diversity is a reflection of multiple coexisting lineages. Investigating the community structure of *B. dolosa* within Patient 1, we found a deeply branched phylogeny with 6 lineages separated by at least 5 lineage-specific mutations (Figure 3.3a). On average, pairs of isolates from this sample differed by 26 mutations, and, of all 406 possible isolate pairs, only one was identical. Thus, even within a single sputum sample, the population is so diverse that full identity between isolates is extremely rare.

In one patient (Patient 5), the *B. dolosa* community had many more mutations than other patients' populations ($P < 0.05$, Grubbs' test for outliers). This excess of mutations is due solely to increased transitions and not transversions, suggesting hypermutation (Figure S2.2a, $P < 0.01$, Grubbs' test). A search of the 199 mutated genes unique to Patient 5's population revealed a single mutation involved in DNA repair: a nonsynonymous mutation at a conserved position in *mutL,* defects of which are known to cause excess transitions[26] (Figure S2.2b). These excess mutations are enriched in synonymous mutations relative to the other patients, further supporting hypermutation ($P < .001$, Figure S2.2c). While hypermutation is a common phenotype in many pathogens, hypothesized to accelerate the evolution of antibiotic resistance[27-30], it has not been previously described in members of the *B. cepacia* complex[31].

For how long have these diverging lineages coexisted? The time to the last common ancestor (LCA) of each non-hypermutating patient's population[32] can be estimated using the number of mutations accumulated since the LCA and the molecular clock previously measured for this outbreak (2.1 SNPs/year[8]). Given the phylogeny of isolates from Patient 1, we calculated the distribution of the number of mutations since the LCA, d_{LCA}, across the population (Figure

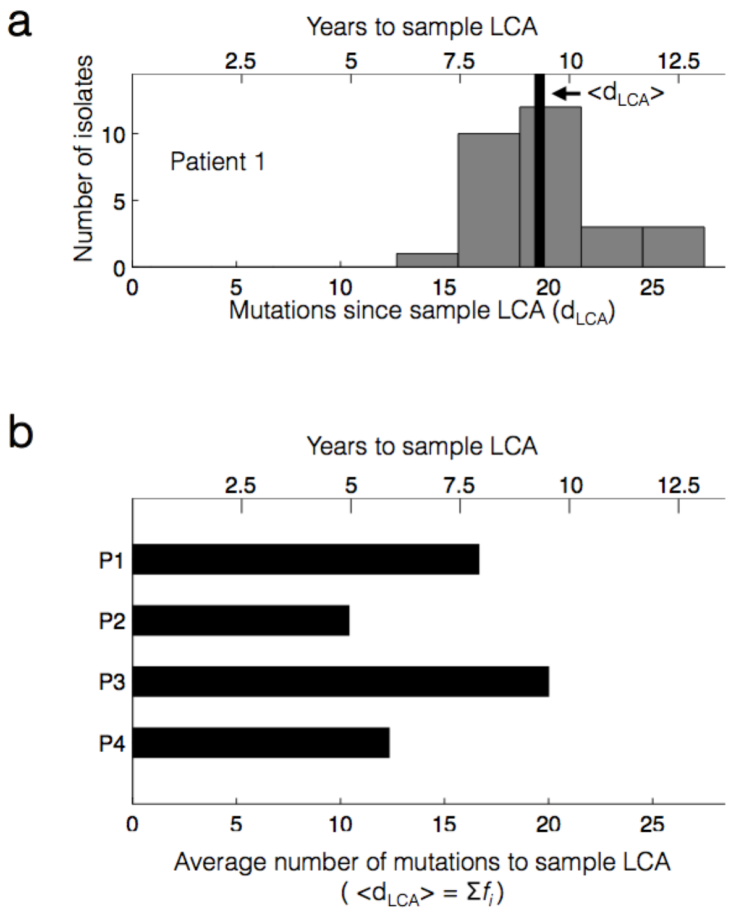

Figure 3.4: Sublineages coexist within a patient for many years after divergence. (**a**) A histogram of the number (d_{LCA}) of single nucleotide mutations found in isolates from Patient 1, relative to their LCA. The black bar indicates the mean value of d_{LCA} across the isolates. (**b**) The value of $<d_{LCA}>$ from the population sequencing data for patients Patients 1 through 4 (Methods). In both panels, the axis at top shows the relationship between d_{LCA} and years to LCA, as calculated via the molecular clock (2.1 SNPs/yr)[8].

3.4a). The mean value of d_{LCA} across isolates, $<d_{LCA}>$, is 19.6 single nucleotide mutations per genome (95% confidence interval, CI = 18.3-20.8), suggesting that the LCA existed 9.3 years ago (CI = 8.7-9.9). This places the LCA of the isolates from this sample slightly earlier than the first *B. dolosa* culture from this patient (7.6 years before sample collection), suggesting that the *B. dolosa* population in Patient 1 has been diverging since, or perhaps before, initial colonization. While the population sequencing approach cannot provide a distribution of d_{LCA}, due to a lack of information regarding linkage between mutations, we can still calculate $<d_{LCA}>$: it is the sum of the polymorphic mutation frequencies (see Supplementary Information 2 for derivation). Using this approach, the estimated time to LCA for Patient 1's population is 7.9 years. This value is slightly lower than calculated from the clonal re-sequencing approach, likely due to mutations left undetected by our conservative polymorphism caller (see Supplementary Information 2 for discussion of error). For Patients 2 and 4, the time to LCA calculated by this population sequencing approach is several years less than the time since first positive culture, suggesting fixation events sometime during these patients' histories (Table S2.1). For all these patients, we estimate that diverging lineages have coexisted in each of these patients for at least 5 years (Figure 3.4b).

To explore the drivers of this long-coexisting diversity, we examined the identity of the evolving genes. Interestingly, we found that within each sample, several *B. dolosa* genes carried 2-4 coexisting polymorphisms (Table S2.2). This clustering is a significant departure from a neutral model given the number of mutations and the distribution of gene lengths (Figure 3.5a, $P < 0.005$ for Patients 1-4; Methods). A similar analysis at the operon-level further identified several operons enriched for polymorphisms (Table S2.3 and Figure S2.3). An enrichment of nonsynonymous mutations in these multi-diverse genes and operons suggests that they are

Figure 3.5: Coexistence of alternative adaptive mutations in the same sample highlights specific genes as drivers of within-host evolution. (**a**) Number of multi-diverse genes observed in samples from Patients 1-5 (P1-P5, blue bars) relative to a null expectation in which diverse sites are randomly distributed across the genome (histogram, 1000 simulations). For P5, the number of multi-diverse genes observed is not significant. (**b**) The canonical signal for selection, dN/dS, across the set of 16 genes and 3 operons showing a multi-diverse signature in at least one patient (P1-P4, 21 genes total, blue) versus dNdS across the set of genes not showing this signature (black). dN/dS >1 indicates positive selection for amino acid change. Error bars indicate 95% CIs. See Online Methods for details on the calculation of dNdS. (**c-d**) Linkage between nearby polymorphisms based on jointly overlapping short reads. Percentages of reads supporting the ancestral genotype, each of the single mutants, and the double mutant are plotted. No reads supporting the double mutant were found (c, n=524; d, n=415; See Figure S2.5 for exception). (e) A network of patients and genes showing a multi-diverse signature at least once in P1-P4. A gene is connected to a patient if it was mutated multiple times (solid line), had a single polymorphic mutation (dashed lined), or single fixed mutation (dotted line) within that patient. Genes closer to the center of the network are mutated in more patients, representing common targets of *in vivo* pathogen selection, while genes connected to single patients may indicate patient-specific adaptation. Genes are labeled with their closest homolog and predicted biological role. The biological role of r*poD* is unclassified because it is recently duplicated in the *B. cepacia* complex[39] (see Supplementary Information 2, Table S2.2).

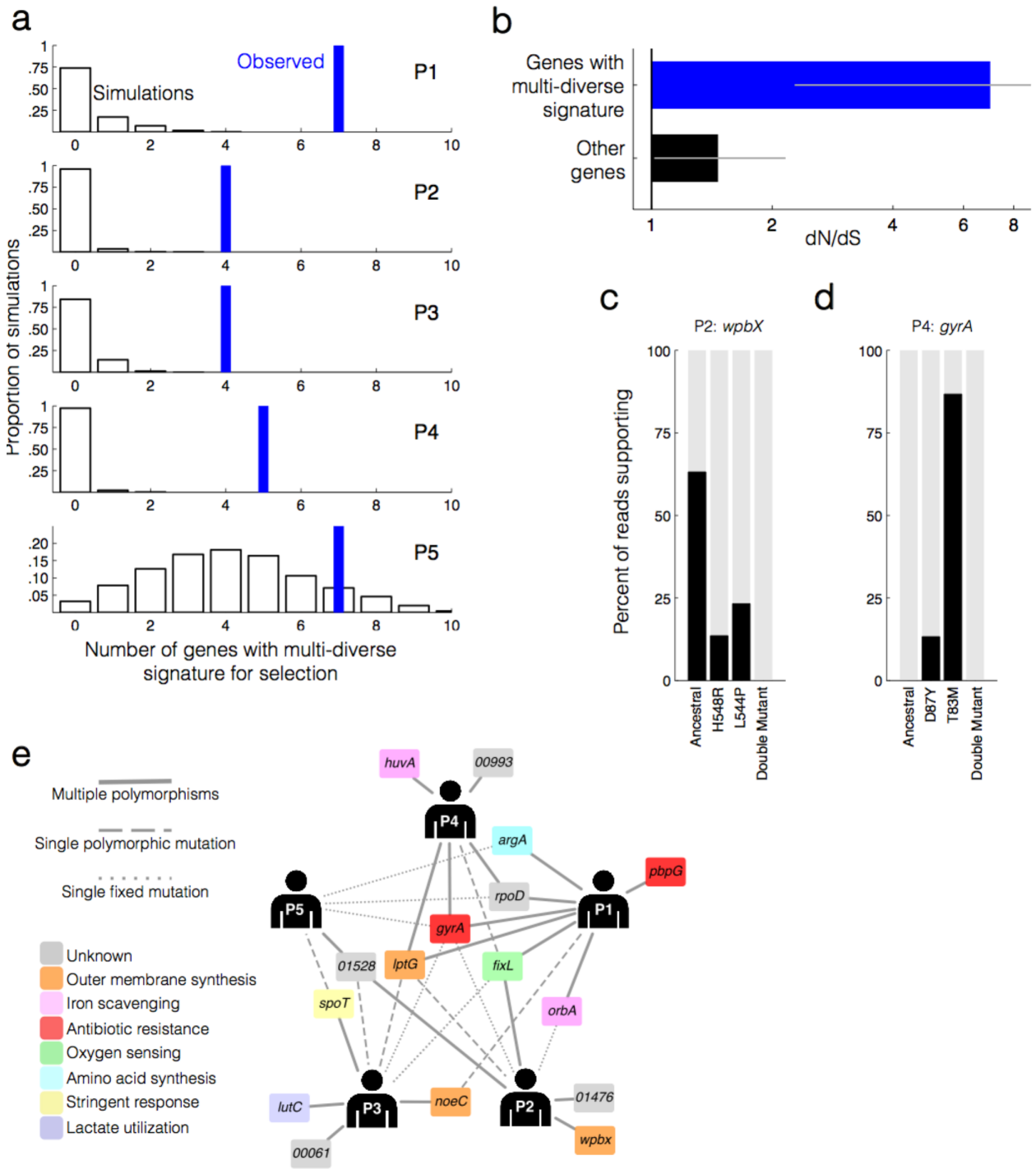

Figure 3.5: Coexistence of alternative adaptive mutations in the same sample highlights specific genes as drivers of within-host evolution (Continued).

drivers of adaptive change *in vivo* (dN/dS = 7.0, CI = 2.3-34.9, Figure 3.5b). Polymorphisms are thus concentrated within genes undergoing adaptive evolution.

To understand why polymorphisms cluster within some genes, we asked if coexisting mutations in the same gene appeared in different lineages or were linked in a double mutant. Examining the single isolate genomes, we found no isolates with doubly mutated genes (Figure S2.4). Similarly, for the population sequencing, in 10 of 11 cases where polymorphic positions are close enough on the genome to be covered by the same short sequencing reads, we did not find reads that contain both variants (Figure 3.5c, Figure S2.5). In some of these cases, the ancestral genotype is completely purged from the population (Figure 3.5d). Thus, diversification is driven by multiple adaptive mutations in the same genes evolving in parallel within individual patients.

These findings provide a new signature of past selective pressures detectable in a single clinical sample; the coexistence of multiple polymorphisms within the same gene in a clinical sample. Sixteen *B. dolosa* genes display this multi-diverse signature, including genes with homologs involved in outer membrane synthesis, antibiotic resistance, iron scavenging, oxygen sensing, amino acid synthesis, lactate utilization, and stress response. Additionally, some genes with less characterized biological roles display a multi-diverse signature, including two transcriptional regulators with unknown targets in *B. dolosa*, an uncharacterized glucoamylase, and two genes that encode hypothetical proteins (Table S2.2). A similar signature for selection is seen in three operons, two involved in lipopolysaccharide transport and one containing a two-component regulatory system with unknown targets (Table S2.3). Selection on many of these elements can be rationalized based on the relevance of their annotated functions to conditions to which the bacteria are exposed in the course of the infection. Yet, further investigation will be

required to understand the potential roles of some of these genes in antibiotic resistance, fitness, and other aspects of pathogenesis.

We found that many of the selective forces acting on the pathogen are the same across patients (Figure 3.5e). Often, genes showing a multi-diverse signature for selection in one patient also carry mutations in other patients ($P < 0.002$, hypergeometric test). A prominent example is *gyrA*, a well-studied target of quinolones, which is mutated in all patient populations. Further support for commonality in mutational trajectories across patients emerges from a significant overlap between this list of 16 multi-diverse genes and 17 genes previously found to be under parallel evolution across a larger group of patients, only one of whom (Patient 2) was included in both studies ($P < 0.001$, hypergeometric test). Thus, the study of a single clinical sample can provide generalizable lists of selective pressures felt within the human body.

Yet, some multi-diverse signatures are patient-specific. A penicillin-binding protein (BDAG_01166, homologous to PBP7) has 3 nonsynonymous mutations in Patient 1, but is not mutated in other patients. Such patient-specific parallel evolution might reflect patient-specific selective pressure or perhaps a fitness benefit dependent upon previously acquired mutations. But these hypotheses are hard to test because the genomic target for a selective force might include more than one gene. For example, four of the five patients' populations have a mutation in a homolog of the histidine kinase *fixL* (BDAG_01161, known to be under strong selection in these infections[8]) while the fifth has a mutation in the corresponding response regulator.

To investigate the stability of these multi-diverse signatures for selection, we collected a second sputum sample 14 days after initial sample collection from Patient 2. Three of the four genes with the multi-diverse signature at day 0 show the same pattern at day 14. The absence of the signature in the fourth gene at the later time point does not reflect a relaxation in selection for

mutant alleles, but rather incomplete detection of genes under selection; this gene also has abundant nonsynonymous mutants at day 14, concentrated at a single nucleotide position (Figure S2.6). These results suggest that the multi-diverse signature for selection is relatively stable and that multiple sample collections per patient can increase the sensitivity of the detection.

Conclusions

Our results reject the dominant lineage model of infection, yet demonstrate that these diversifying bacteria adapt under the pressure of natural selection. These observations are consistent with clonal interference: in large asexual populations, multiple beneficial mutations emerge and compete, impeding the ability of these lineages to reach fixation[33-35]. In addition to large population size (10^8 cells/mL sputum), the branched structure of the airways may further hinder the capacity of any adaptive lineage to dominate and fix, and the immune system or niche-specific adaptations might directly promote diversity. Diversified by any of these means, lineages may then continue to evolve in parallel against common selective forces.

As *B. dolosa* adapts to the airways of people with cystic fibrosis, mutations lead to diversification rather than fixation and replacement. Though it is possible that adaptive mutations will lead to fixation more frequently in other infections, there is evidence that, at least in long-term colonization, diversity might be common[14,36-38]. This long-term coexistence of diverse lineages records the genomic history of selection on the pathogen within its host. The ability to rapidly read off within-patient evolutionary history from the genotypic diversity within a single clinical sample may greatly accelerate the ability to survey selective pressures acting on bacterial pathogens *in vivo* – shifting from an epidemic level investigation to a single-patient paradigm.

Methods

Study cohort and sample collection

An epidemic clone of *B. dolosa* infected and colonized 39 individuals with cystic fibrosis in the Boston area over a 20-year period[21]. We studied *B. dolosa* intrapatient diversity in 5 surviving individuals still infected with *B. dolosa*. All subjects were male, had homozygous ΔF508 mutations, had not received lung transplants, were between 21 and 35 years of age, and had been colonized for between 7 and 10 years at the time of sample collection (see Table S2.1). Longitudinal microbial isolates from Patient 2 were also included in a previous retrospective study (patient J in reference 8).

For Patient 1, both the colony re-sequencing and deep population sequencing approaches were performed on a single sputum sample (P1). For Patient 2, population deep sequencing was performed on each of two sputum samples (P2 and P2T), collected 14 days apart. Between collections, Patient 2 was treated for a pulmonary exacerbation, including a change in antibiotic regimen, but his condition did not improve and *B. dolosa* density did not decrease. For Patients 3-5, population sequencing was performed on a single sputum sample from each patient (P3-P5).

Expectorated sputum samples were collected at Boston Children's Hospital after written informed consent was obtained under protocols approved by the Institutional Review Boards at Boston Children's Hospital and Harvard Medical School. Samples were liquefied with dithiothreitol[40] and stored at -80°C in 20% glycerol. *B. dolosa* was cultured from frozen samples. For population sequencing, a plate with 5,000 to 30,000 small colonies was chosen from a serial dilution. See Supplementary Information 2 for more details on sample preparation.

Illumina sequencing

Genomic DNA was extracted using MoBio UltraClean Microbial DNA Isolation Kit per the manufacturer's instructions. Genomic libraries were constructed and barcoded using the Illumina-compatible Epicentre Nextera DNA Sample Prep Kit and following manufacturer's instructions (PCR amplification in the Nextera preparation does not introduce false positive polymorphisms, see Supplementary Information 2). Genomic libraries were sequenced on the Illumina HiSeq 2000 by Partners HealthCare Center for Personalized Genetic Medicine. Individual colonies were sequenced using single-end, 50 base-pair (bp) reads and pooled samples were sequenced using paired-end, 50bp reads. Reads were aligned to the *B. dolosa* draft genome AU0158 (GenBank accession number AAKY00000000, see URLs), belonging to an isolate recovered from patient zero of the outbreak. AU0158 consists of 233 contigs on 3 scaffolds (*B. dolosa* has 3 chromosomes). Standard approaches were used for read filtering and alignment (Supplementary Information 2). See Table S2.4 for coverage statistics.

Mutation identification, colony re-sequencing

An outgroup of 3 outbreak strains (A-0-0, G-9-8, and N-12-6d-$, previously sequenced[8]) was included in the analysis to identify mutations fixed among the 29 isolates from P1. We considered genomic positions at which at least one pair of isolates was discordant on the called base and both members of the pair had FQ scores less than -40 (FQ scores are produced by SAMtools[41]; lower values indicate agreement amongst reads). Genomic positions for which multiple isolates had multiple calls per isolate were discarded (likely duplication not represented in the reference). A best call was forced for each isolate (Table S2.5) and the list of concatenated

SNPs was inputted into the dnapars program in PHYLIP v3.69[42]. The resulting phylogeny was visualized the tree using Figtree (Figure 3.3b).

Mutation identification, deep population sequencing

Fixed mutations within each patient's population were called using the same procedure as individual isolates, with a stricter quality score threshold (FQ < -282). Custom MATLAB scripts and SAMtools-produced pileup files were used to summarize all calls and their related quality scores at each genomic position (e.g. base quality, mapping quality, tail distance; see Supplementary Information 2). Using the isogenic control, multiple isolates from Patient 1, and an interactive MATLAB environment that enabled investigation of the raw data, we developed a set of filters to identify true-positive polymorphic positions with minor allele frequency above 3% (Table S2.6). Thresholds were chosen to minimize false positives. See Supplementary Information 2 and Figures S2.7-S2.8 for description of filters and sensitivity analysis.

Estimation of $\langle d_{LCA} \rangle$

For the colony-based approach, d_{LCA} was calculated for each isolate as the number of mutations received by that isolate normalized by the size of the callable genome. For this approach, the callable genome is the set of genomic positions with FQ score < -40. The confidence interval for $\langle d_{LCA} \rangle$ presented for this approach is calculated according to a Poisson distribution. For the pool-based approach, $\langle d_{LCA} \rangle$ was calculated as the sum of the mutation frequencies at each polymorphic position called within that population, normalized by the size of the callable genome (see Supplementary Information 2). For the pool-based approach, we define the callable genome as the set of positions that met the chosen thresholds for coverage, average

base quality, average mapping quality, and average tail distance for each strand, irrespective of nucleotide call. See Figure S2.6b and the Supplementary Information 2 for a discussion of sources of error in estimating $<d_{LCA}>$ and time to LCA.

Detection of parallel evolution within patients

We define genes with a multi-diverse signature of selection as genes for which within the same sputum sample there were multiple polymorphisms and multiple polymorphisms per 2000bp (to account for the fact that long genes are more likely to be mutated multiple times by chance). To determine whether the number of genes showing this signature was a significant departure from what expected in a neutral model, we performed for each sputum sample 1000 simulations in which we randomly shuffle the polymorphisms found across the callable genome, and calculate how many genes show the signature of selection (Figure S2.5a).
This analysis was repeated at the operon and pathway levels, using the free version FgenesB to identify operons and subsystem annotations provided by SEED[43] as pathways (see Figure S2.3). As in the gene analysis, we considered operons and pathways to have a signature for selection if they had both multiple polymorphisms and multiple polymorphisms per 2000 nucleotides with the same patient.

dN/dS

Mutations were classified as nonsynonymous (N) or synonymous (S) according to annotations provided in genbank file. For open reading frames in draft genome without a provided frame, we used BLAST and RefSeq to identify the most likely reading frame in the neighborhood of the found mutations. For each dNdS calculation, we used the particular

spectrum of mutations observed to calculate the expected N/S (e.g. A->C mutations are 10.6 times more likely to cause an N than G->A mutations)**.** The observed value of N/S was divided by this expectation to get dN/dS. Confidence intervals and p-values were calculated according to binomial sampling. The dNdS reported Figure 3.5b groups together the mutations found in genes and operons under selection; the same calculation for only genes gives a dN/dS of 5.9 (95% CI = 1.9-29.6).

Parallel evolution across patients

We used the hypergeometric distribution to assess the significance of overlap between gene sets. Of 225 *B. dolosa* genes mutated in P1-P4, only 16 showed the multi-diverse signature for selection within patients and only 29 genes were mutated in multiple of these patients (fixed or polymorphic), yet 7 genes are in common between these lists (P=.0015) Similarly, 13 of these 225 genes were also found on a list of 17 genes evolved in parallel across patients in a previous study[8]. These 13 genes were enriched in the 16 genes under selection in in this study (5 gene overlap, P=.0009). When this analysis was repeated without mutations from P2 (Patient 2 also included in retrospective study), 11 of the 189 mutated genes were found in the previous study and 13 genes show a multi-diverse signature for selection. The overlap between these lists of 11 and 13 genes is smaller but still significant (4 genes; P=.0035).

Contributions

TDL, AJM, GPP and RK designed the study. AJM and TRM collected clinical samples. KBF, TRM, AJM and GPP conducted chart review and provided medical information. TDL performed experiments. TDL, IY and RK wrote the sequence analysis scripts. TDL and RK analyzed the data. TDL, AJM, GPP and RK interpreted the results and wrote the paper.

Acknowledgements

We are grateful to Jean-Baptiste Michel and members of the Kishony laboratory for insightful discussions and support, to the team at the Partners HealthCare Center for Personalized Genetic Medicine (PCPGM) for Illumina sequencing, to L. Williams and A. Palmer for discussions and technical assistance, and to Y. Gerardin, J. Meyer, L. Stone and R. Ward for their comments on the manuscript. TDL and GPP were supported in part by grants from the Cystic Fibrosis Foundation (LIEBER12H0 to TDL and PRIEBE1310 to GPP). This work was funded in part by the US National Institutes of Health (GM081617 to RK), the New England Regional Center of Excellence for Biodefense and Emerging Infectious Diseases (NERCE; U54 AI057159 to RK) and Hoffman-LaRoche.

References

1. Mwangi, M.M. *et al*. Tracking the in vivo evolution of multidrug resistance in *Staphylococcus aureus* by whole-genome sequencing. *Proc Natl Acad Sci U S A* **104**, 9451-6 (2007).

2. Comas, I. *et al*. Whole-genome sequencing of rifampicin-resistant *Mycobacterium tuberculosis* strains identifies compensatory mutations in RNA polymerase genes. *Nat Genet* **44**, 106-10 (2012).

3. Ford, C.B. *et al*. Use of whole genome sequencing to estimate the mutation rate of *Mycobacterium tuberculosis* during latent infection. *Nat Genet* **43**, 482-6 (2011).

4. Kennemann, L. *et al*. *Helicobacter pylori* genome evolution during human infection. *Proc Natl Acad Sci U S A* **108**, 5033-8 (2011).

5. Young, B.C. *et al*. Evolutionary dynamics of *Staphylococcus aureus* during progression from carriage to disease. *Proc Natl Acad Sci U S A* **109**, 4550-5 (2012).

6. Huse, H.K. *et al*. Parallel evolution in *Pseudomonas aeruginosa* over 39,000 generations in vivo. *MBio* **1**(2010).

7. Snitkin, E.S. *et al*. Tracking a Hospital Outbreak of Carbapenem-Resistant *Klebsiella pneumoniae* with Whole-Genome Sequencing. *Science Translational Medicine* **4**, 148ra116-148ra116 (2012).

8. Lieberman, T.D. *et al*. Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nat Genet* **43**, 1275-80 (2011).

9. Didelot, X. *et al*. Genomic evolution and transmission of *Helicobacter pylori* in two South African families. *Proceedings of the National Academy of Sciences* (2013).

10. Wilson, D.J. Insights from genomics into bacterial pathogen populations. *PLoS Pathog* **8**, e1002874 (2012).

11. Workentine, M. & Surette, M.G. Complex *Pseudomonas* Population Structure in Cystic Fibrosis Airway Infections. *American journal of respiratory and critical care medicine* **183**, 1581-1583 (2011).

12. Nguyen, D. & Singh, P.K. Evolving stealth: genetic adaptation of *Pseudomonas aeruginosa* during cystic fibrosis infections. *Proc Natl Acad Sci U S A* **103**, 8305-6 (2006).

13. Chung, J.C. *et al*. Genomic variation among contemporary *Pseudomonas aeruginosa* isolates from chronically infected cystic fibrosis patients. *J Bacteriol* **194**, 4857-66 (2012).

14. Workentine, M.L. *et al*. Phenotypic Heterogeneity of *Pseudomonas aeruginosa* Populations in a Cystic Fibrosis Patient. *PloS one* **8**, e60225 (2013).

15. Foweraker, J.E., Laughton, C.R., Brown, D.F. & Bilton, D. Phenotypic variability of *Pseudomonas aeruginosa* in sputa from patients with acute infective exacerbation of cystic fibrosis and its impact on the validity of antimicrobial susceptibility testing. *J Antimicrob Chemother* **55**, 921-7 (2005).

16. Sun, G. *et al*. Dynamic population changes in *Mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients. *J Infect Dis* **206**, 1724-33 (2012).

17. Harris, S.R. *et al*. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *The Lancet infectious diseases* (2012).

18. Walker, T.M. *et al*. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *The Lancet infectious diseases* (2012).

19. Hansen, S.K. *et al*. Evolution and diversification of *Pseudomonas aeruginosa* in the paranasal sinuses of cystic fibrosis children have implications for chronic lung infection. *The ISME journal* (2011).

20. Vermis, K. *et al*. Proposal to accommodate *Burkholderia cepacia* genomovar VI as *Burkholderia dolosa* sp. nov. *International journal of systematic and evolutionary microbiology* **54**, 689-691 (2004).

21. Kalish, L.A. *et al*. Impact of *Burkholderia dolosa* on lung function and survival in cystic fibrosis. *Am J Respir Crit Care Med* **173**, 421-5 (2006).

22. Cibulskis, K. *et al*. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* **31**, 213-219 (2013).

23. Barrick, J.E. & Lenski, R.E. Genome-wide mutational diversity in an evolving population of *Escherichia coli*. *Cold Spring Harb Symp Quant Biol* **74**, 119-29 (2009).

24. Pickrell, J.K., Gilad, Y. & Pritchard, J.K. Comment on "Widespread RNA and DNA sequence differences in the human transcriptome". *Science* **335**, 1302; author reply 1302 (2012).

25. Nakamura, K. *et al*. Sequence-specific error profile of Illumina sequencers. *Nucleic acids research* **39**, e90-e90 (2011).

26. Oliver, A. & Mena, A. Bacterial hypermutation in cystic fibrosis, not only for antibiotic resistance. *Clin Microbiol Infect* **16**, 798-808 (2010).

27. Oliver, A., Canton, R., Campo, P., Baquero, F. & Blazquez, J. High frequency of hypermutable *Pseudomonas aeruginosa* in cystic fibrosis lung infection. *Science* **288**, 1251-4 (2000).

28. Jolivet-Gougeon, A. *et al*. Bacterial hypermutation: clinical implications. *J Med Microbiol* **60**, 563-73 (2011).

29. Hoboth, C. *et al*. Dynamics of adaptive microevolution of hypermutable *Pseudomonas aeruginosa* during chronic pulmonary infection in patients with cystic fibrosis. *J Infect Dis* **200**, 118-30 (2009).

30. Marvig, R.L., Johansen, H.K., Molin, S. & Jelsbak, L. Genome Analysis of a Transmissible Lineage of *Pseudomonas aeruginosa* Reveals Pathoadaptive Mutations and Distinct Evolutionary Paths of Hypermutators. *PLoS genetics* **9**, e1003741 (2013).

31. Pope, C.F., Gillespie, S.H., Moore, J.E. & McHugh, T.D. Approaches to measure the fitness of *Burkholderia cepacia* complex isolates. *J Med Microbiol* **59**, 679-86 (2010).

32. Kingman, J.F.C. On the genealogy of large populations. *Journal of Applied Probability*, 27-43 (1982).

33. Fogle, C.A., Nagle, J.L. & Desai, M.M. Clonal interference, multiple mutations and adaptation in large asexual populations. *Genetics* **180**, 2163-73 (2008).

34. Gerrish, P.J. & Lenski, R.E. The fate of competing beneficial mutations in an asexual population. *Genetica* **102-103**, 127-44 (1998).

35. Hegreness, M., Shoresh, N., Hartl, D. & Kishony, R. An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science* **311**, 1615-7 (2006).

36. Mowat, E. *et al*. *Pseudomonas aeruginosa* population diversity and turnover in cystic fibrosis chronic infections. *Am J Respir Crit Care Med* **183**, 1674-9 (2011).

37. Schloissnig, S. *et al*. Genomic variation landscape of the human gut microbiome. *Nature* **493**, 45-50 (2013).

38. Ashish, A. *et al*. Extensive diversification is a common feature of *Pseudomonas aeruginosa* populations during respiratory infections in cystic fibrosis. *Journal of Cystic Fibrosis* (2013).

39. Menard, A., de Los Santos, P.E., Graindorge, A. & Cournoyer, B. Architecture of *Burkholderia cepacia* complex sigma70 gene family: evidence of alternative primary and clade-specific factors, and genomic instability. *BMC Genomics* **8**, 308 (2007)

40. Guss, A.M. *et al*. Phylogenetic and metabolic diversity of bacteria associated with cystic fibrosis. *ISME J* **5**, 20-9 (2011).

41. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

42. 42.Felsenstein, J. PHYLIP-Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166 (1989).

43. Aziz, RK *et al*. The RAST Server: rapid annotations using subsystems technology. *BMC genomics* **9**, 75 (2008).

Chapter 4:

Extension of genomic approaches to study intraspecies diversification across space

The finding of sustained intraspecies microbial diversity in Chapter 3 has prompted multiple new collaborations. Here, I describe my progress in two of these directions, both of which investigate the role of spatial differences in the maintenance of genomic diversity. In the first study, we are characterizing *Mycobacterium tuberculosis* diversity within and between organs, from many individuals co-infected with HIV. In the second, we are surveying intraspecies diversity of *Stenotrophomonas maltophilia* in 32 tissue samples taken from the same explanted cystic fibrosis lung. These studies will illuminate the mechanistic basis of diversification of these infections, point to genes important to survival of these pathogens in different niches within the body, and demonstrate that the analytical approaches developed in Chapter 2 and 3 are widely applicable methods for gaining insight into microbial pathogens.

Introduction

Our finding that genotypically diverse descendants of an initial colonizing strain can co-exist within the airways of a CF patient for long periods of time raises an obvious next question—do these different genotypes co-exist in the same microenvironment, or do they colonize distinct regions of the airway? A similar, and perhaps orthogonal, unknown is if these genotypes have distinct ecological niches. Assessing intraspecies diversity over space can address both the spatial and ecological components of within-patient variation.

Towards this end, I have begun working on two collaborative projects to explore intraspecies diversity over space— focusing on *Mycobacterium tuberculosis* diversity across the many body sites of many patients most-mortem and *Stenotrophomonas maltophilia* diversity at finer-resolution within an explanted cystic fibrosis lung. While these two studies (described in more detail below), focus on distinct infections and at different levels of spatial resolution, there are many commonalities in the approach and questions addressed.

In both studies, we are sequencing the whole genomes of multiple bacterial isolates from each site and identifying mutations relative to a reference genome. Ongoing phylogenetic analyses will show the ancestral relationship of strains at different body sites, providing insights into transmission routes and migration rates across body sites. The degree of diversity across and within sites will have implications for diagnosis and testing, and will inform better models of within-patient evolution and the emergence of antibiotic resistance. Integration of location, phenotype, and genotype information will illuminate the mechanistic basis of the observed genomic differentiation within patients.

As in the previous chapters, we are also searching for evidence of parallel evolution across and within patients in order to identify the selective forces felt by bacteria within the

body. We will use the same phylogenetic approaches performed in Chapter 2 to identify nucleotides that evolved in parallel. We will look for parallel evolution within a body site, which may suggest site-specific selective forces. By integrating these genomic signals with site-specific physiological information, we may be able to better understand the role of these genes to bacterial survival in the body. For example, in the explant lung, co-localization of another microbial pathogen (surveyed using 16S profiling) with *S. maltophilia* genotypes harboring mutations in a particular gene might suggest that this gene modulates interspecies interactions.

Within-patient diversity and evolution of *Mycobacterium tuberculosis*

We are involved in a collaboration with Theodore Cohen, an epidemiologist and associate professor at the Harvard School of Public Health, and others to understand the diversity present in individual *M. tuberculosis* infections and across organs. This ongoing study is based in the province of Kwazulu-Natal, South Africa (KZN), where patient-enrollment, sample-collection, culturing of *M. tuberculosis*, antibiotic-resistance profiling, and DNA extraction are being performed. In January 2012, I had the opportunity to travel with Dr. Cohen to KZN and meet some of our collaborators, including Douglas Wilson, the Head of the Department of Medicine at Edendale Hosptial, and Professor Preshnie Moodley at the University KZN.

We are comparing *M. tuberculosis* diversity across the body of many individuals, taking multiple biopsies from each individual post-mortem in a limited autopsy. We joined Dr. Cohen after he and our other collaborators had already begun on this exciting direction, but before sample collection had begun. In this way, we were able to aid in study design to address both epidemiological and evolutionary questions about *M. tuberculosis* diversity.

Initial sample collection is now complete, with 100 autopsies performed. We are focusing on co-infections with HIV, in which *M. tuberculosis* often spreads from the lung to other parts of the body. To have the best chance of capturing interstrain and intrastrain diversity, we only sampled recently deceased individuals who had been on antitubercular therapy for fewer than 5 days. We sampled from the respiratory tract (endotracheal aspirate), lung tissue, liver, spleen, lymph node, and pleural fluid of each patient, taking multiple biopsies from each location to maximize the probability of positive cultures. Biopsies from the lung were cultured independently to enable comparison of intra-organ and inter-organ diversity. From other sites, biopsies were pooled and cultured by site for practical considerations.

Microbial culturing is ongoing for most of these autopsies, and we will analyze the microbial diversity for any patient with multiple culture-positive sites (~70%). The DNA from each of multiple colonies from culture (up to 15 per site) is being extracted in in the Moodley lab and processed in the Kishony lab for sequencing. We have received hundreds of DNA samples already and sequenced the DNA 98 colonies. Analysis of these samples is ongoing. Initial analysis suggests that single colonies are formed by genetically heterogeneous bacteria, so we will need to employ the approaches developed in Chapter 3 to identify *de novo* mutations.

Intraspecies diversity across an explant cystic fibrosis lung

Following our preliminary findings of vast coexisting *Burkholderia dolosa* genomic diversity, together Gregory Priebe, Alexander McAdam, Roy Kishony, pathologist Sara Vargas, and infectious disease fellow Dr. Kelly Flett initiated a project at Children's Hospital to sample microbial diversity across an explanted CF lung. Patients with CF occasionally get lung transplants, and researchers can sometimes, with the patient or parent's consent, get access to the

explant lung for study. We devised a plan for sampling many sites from each lung using nearly sterile conditions and working within the framework of a standard pathology report. This project was then joined by Hattie Chung who is now leading the analysis in the Kishony lab.

While some sampling of microbial diversity has been performed across lungs these studies have been limited for the most part to 16S-profiling of interspecies diversity and have focused on the coarse spatial resolution lobes[1-3]. One notable exception used sequencing of loci implicated in antimicrobial resistance to show that *M. tuberculosis* acquires antibiotic resistance independently at different loci within the lung[4]. To the best of my knowledge, no similar study has been conducted with in any other infections or for any organism at the genomic level.

It took over two years from receiving initial IRB to receive our first explanted lung. In October 2012, we sampled 31 tissue sites, each about a cubic centimeter, from the explanted lungs of a patient infected primarily with *Stenotrophomonas maltophilia*. We cut cross-sectional slices through the lung, sampling from many sites underneath each clean cut (**Figure 4.1**). Sampling sites were chosen to represent different locations and pathologies, including large airways, lymph nodes, regions of intense scarring, and healthier tissue. From each site, an adjacent tissue sample was sampled for histological staining and pathology reports. Each tissue was mechanically homogenized in a special grinding conical tube and frozen in 15% glycerol.

We are now analyzing the microbial diversity in this lung, sequencing 24 isolates of *S. maltophilia* from each of 26 sites that grew high densities of this species. We have also used the population-sequencing approach developed in Chapter 3 on each sample, but found that colony re-sequencing—which is now more affordable—will provide more useful information. Preliminary analysis suggests significant *S, maltophilia* diversity within each site and significant differences between sites.

Figure 4.1. Sampling of explanted CF lung. We performed sampling procedure of the explanted lung in October 2012. Cross sectional cuts were made through the lung, and samples were taken from the newly exposed tissue. Two discarded cross sectional pieces are seen in this photograph. From each location, two adjacent samples were taken—one for histology and the other for microbiology. Each cut was made with a fresh blade, and each sample was taken using a new sterile scalpel and forceps. Some cross-contamination may have occurred during slicing; to limit this affect the sample for microbiology was taken from underneath the sample for histology.

References

1. Willner, D. *et al*. Spatial distribution of microbial communities in the cystic fibrosis lung. *ISME J* **6**, 471-4 (2012).

2. Goddard, A.F. *et al*. Direct sampling of cystic fibrosis lungs indicates that DNA-based analyses of upper-airway specimens can misrepresent lung microbiota. *Proceedings of the National Academy of Sciences* **109**, 13769-13774 (2012).

3. Erb-Downward, J.R. *et al*. Analysis of the lung microbiome in the "healthy" smoker and in COPD. *PLoS One* **6**, e16384 (2011).

4. Kaplan, G. *et al*. Mycobacterium tuberculosis Growth at the Cavity Surface: a Microenvironment with Failed Immunity. *Infection and Immunity* **71**, 7099-7108 (2003).

Chapter 5:

Concluding Remarks

The *Burkholderia dolosa* outbreak studied here has proved to be a tractable model system for understanding bacterial evolution *in vivo* and has enabled the development of a powerful analytical toolbox. Evidence of parallel evolution and co-existing diversity in other pathogens (reviewed in Chapter 1) suggests that the approaches developed here may be widely applicable.

The applicability of these approaches to acute infections remains to be seen. While antibiotic resistance can emerge during acute infections, adaptation against other selective pressures faced in the human body has not been shown on this time scale. Should such adaptation occur, it is likely that multiple lineages will evolve in parallel and the single-sample approach developed in Chapter 3 will identify genes under selection.

The ability to rapidly identify challenges to survival for each pathogen *in vivo*, and thus potential therapeutic directions, will become of increasing importance as our supply of antibiotics continues to dwindle. Furthermore, the potential for the suggested therapies to be narrow-spectrum may provide an advantage, as broad-spectrum antibiotics can both disrupt our natural flora in dangerous ways and increase the frequency of resistant bacteria. However, an important challenge remaining is the development of a roadmap for turning knowledge of selective pressures into therapeutic directions. We are currently involved in a collaboration to investigate the role of the histidine kinase *fixL*, the most mutated gene across our studies, and oxygen-dependent regulation in *B. dolosa* pathogenesis with hopes of creating new clinical directions. Similar studies are going for *P. aeruginosa* and other pathogens.

The approaches presented here might also be applied to understanding our commensal flora. Each person's microbiomes contains approximately 10^{20} nucleotides of DNA (~10^{14} bacteria, each with ~10^{6} nucleotides of DNA), and every nucleotide has a probability of about 10^{-10} of mutating during each replication event. Therefore, every turnover of our microbiome contains about 10^{10} new mutations upon which selection can act. While the overwhelming majority of these mutations will be deleterious, any non-deleterious mutation will enable the tracking of commensals within or between body compartments. Furthermore, it is not unreasonable to think that members of our microbiome adapt via mutation against challenges presented by the particular host or microbial community. The identification of such adaptive evolutionary events will add context to species-level microbiome differences and be more straightforward to interpret.

Technical and analytical innovations in the coming years will further accelerate our ability to track within-patient evolution and to translate lists of mutations into biological understanding. Improvements in single-cell sequencing will facilitate the comparison of genomes directly from clinical samples, without any biases introduced by culturing. Increases in scale enabled by continuing declines in the cost of sequencing will permit the inference of mutational order in parallel lineages and may suggest epistatic interactions between mutations. Improvements in the gathering and interpreting of clinical metadata will enable the systematic association of evolutionary change with patient outcome. Carefully designed studies leveraging these and other innovations will address outstanding questions, including the drivers of within-patient diversification and the extent to which adaptation is patient-specific.

Appendix 1:

Supplemental Materials for Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes (Chapter 2)

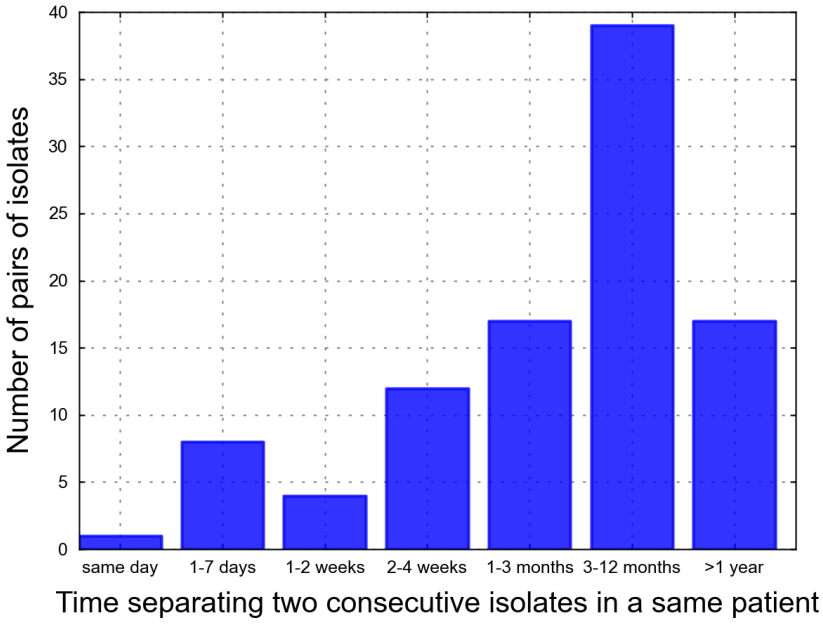

Figure S1.1. Distribution of time separating two consecutive isolates from the same patient. We plot the number of pairs of isolates taken from the same patient within the time windows shown on the x-axis. On average, two consecutive isolates were separated by 3.5 months, but in 9 cases isolates were recovered during the same week.

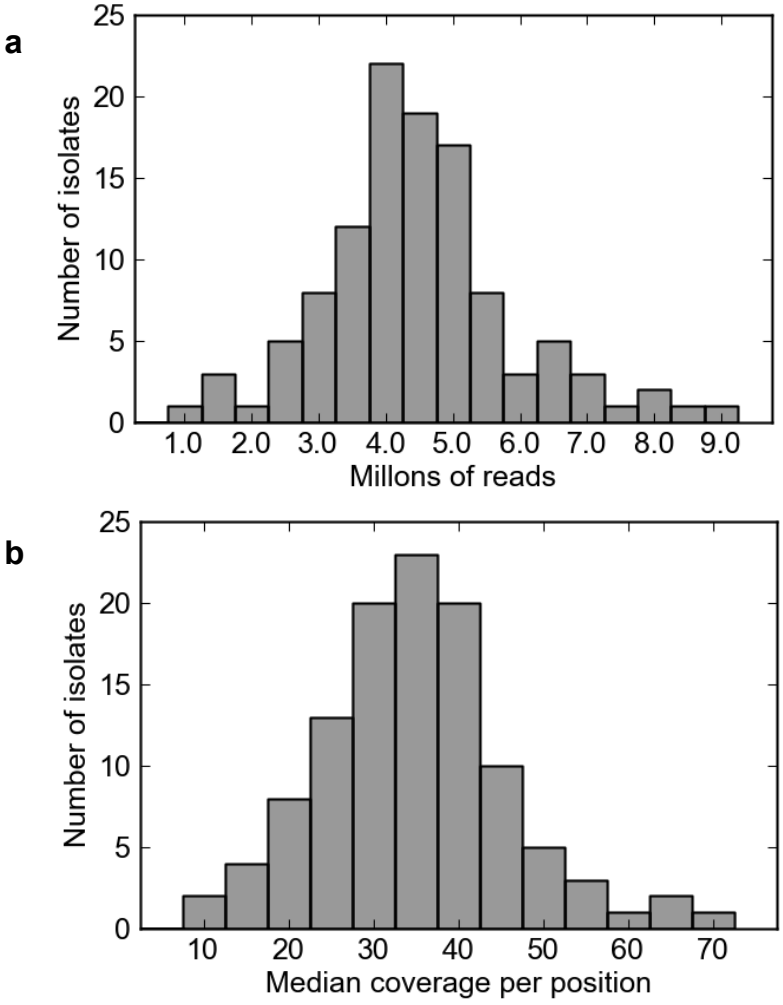

Figure S1.2. Whole genome sequencing of 112 *B. dolosa* isolates at 37x depth. a, Distribution of the number of 75-nucleotide single end reads obtained per bacterial isolate. On average, 4.6 million reads were generated for each isolate. **b,** Distribution of median read depth per position for each isolate. On average, genomes were read to a depth of 37x, providing high quality reads for 93% of the genome.

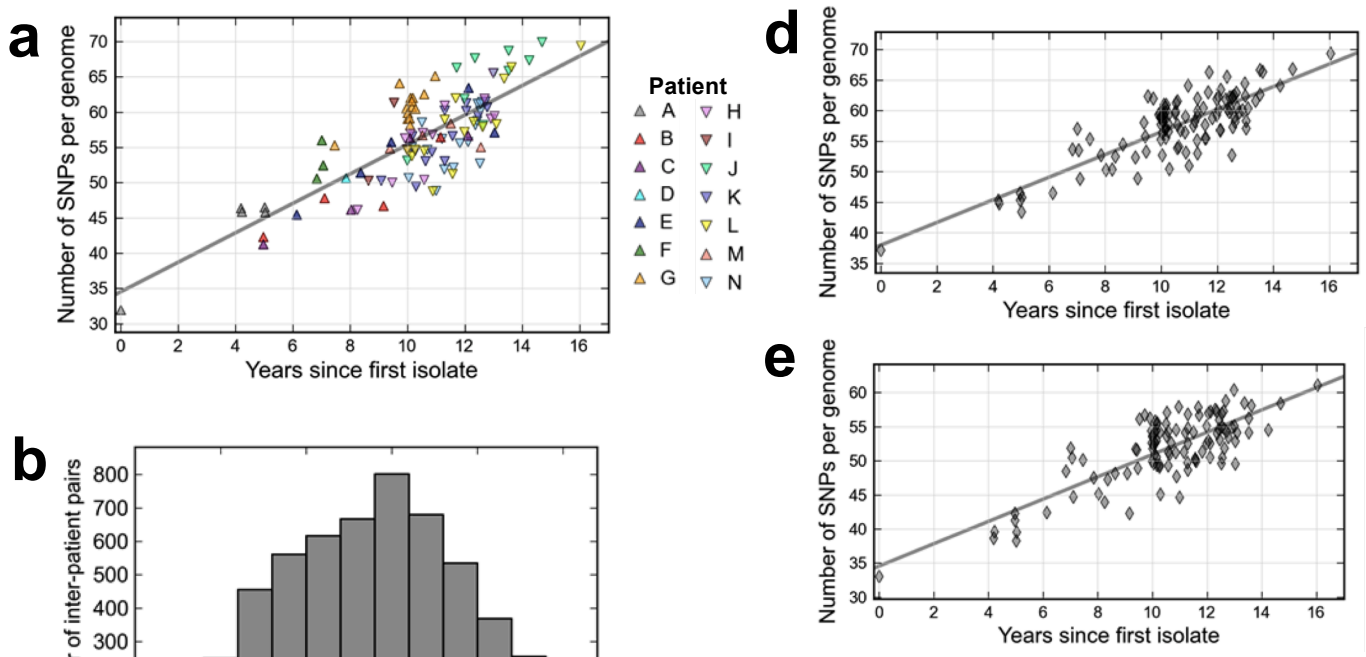

Figure S1.3. Genetic distance between bacterial isolates and rate of evolution. a, The rate at which bacterial mutations accumulate is consistent across patients. We plot the genetic distance of sequenced strains to the outgroup as a function of the time of isolation (years since first isolate in this study). Isolates from the same patient are represented in the same color (see legend). Within patients, isolates accumulate mutations at a rate similar to that observed for all 112 isolates. **Isolates recovered from the same patient are genetically closer.** Distributions of SNP between pairs of isolates taken from different patients **(B)** and from the same patient **(c).** On average, 28 SNPs separate two isolates from different patients; isolates from the same patient differ by an average 9.5 SNPs. **Mutations on genes which are not found under selection accumulate linearly with time. d** We exclude from the calculation of genetic distance those 17 genes under strong selection (which received 3 mutations or more overall). Mutations accumulate linearly with time (Pearson correlation: .83, fixation rate: 1.9 mutations per year). **e,** We exclude from the calculation of genetic distance the 45 genes which received more than one mutation. Mutations accumulate linearly with time (Pearson correlation: .79, fixation rate: 1.6 mutations per year).

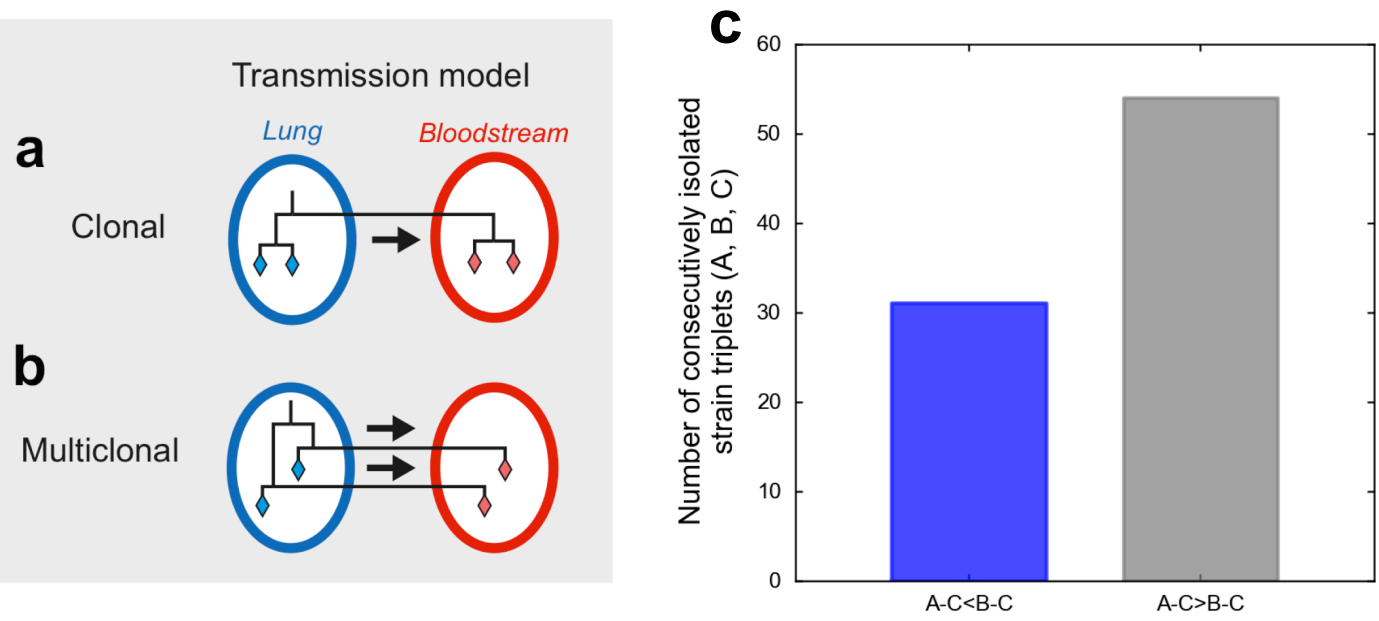

Figure S1.4. Within-patient bacterial populations. a, b, Models of transmission from the lungs to the blood during bacteremia. Monoclonal transmission (**a**): a single clone from the lung is passed to the bloodstream (single arrow). Multiclonal transmission (**b**): multiple clones are transmitted from the lung to the blood, during one or several events (several arrows). **c, Phylogenetic analysis provides indirect evidence that the *B. dolosa* population in the CF lung is polymorphic.** We considered all the triplets of isolates (A,B,C) recovered successively from a patient. For each triplet, we asked whether C is genetically closer to A than to B (A-C < B-C), based on the phylogenetic tree in Figure 2.2b. A positive answer is not expected if the lung was a homogeneous environment for bacteria. We represent in gray the number of triplets that are in line with a homogeneous lung and in blue the number of triplets that contradict this expectation. We find discordance between phylogeny and time in 37% of cases (31 / 85 triplets). This supports a polymorphic model of the lung, where several distinct genotypes coexist.

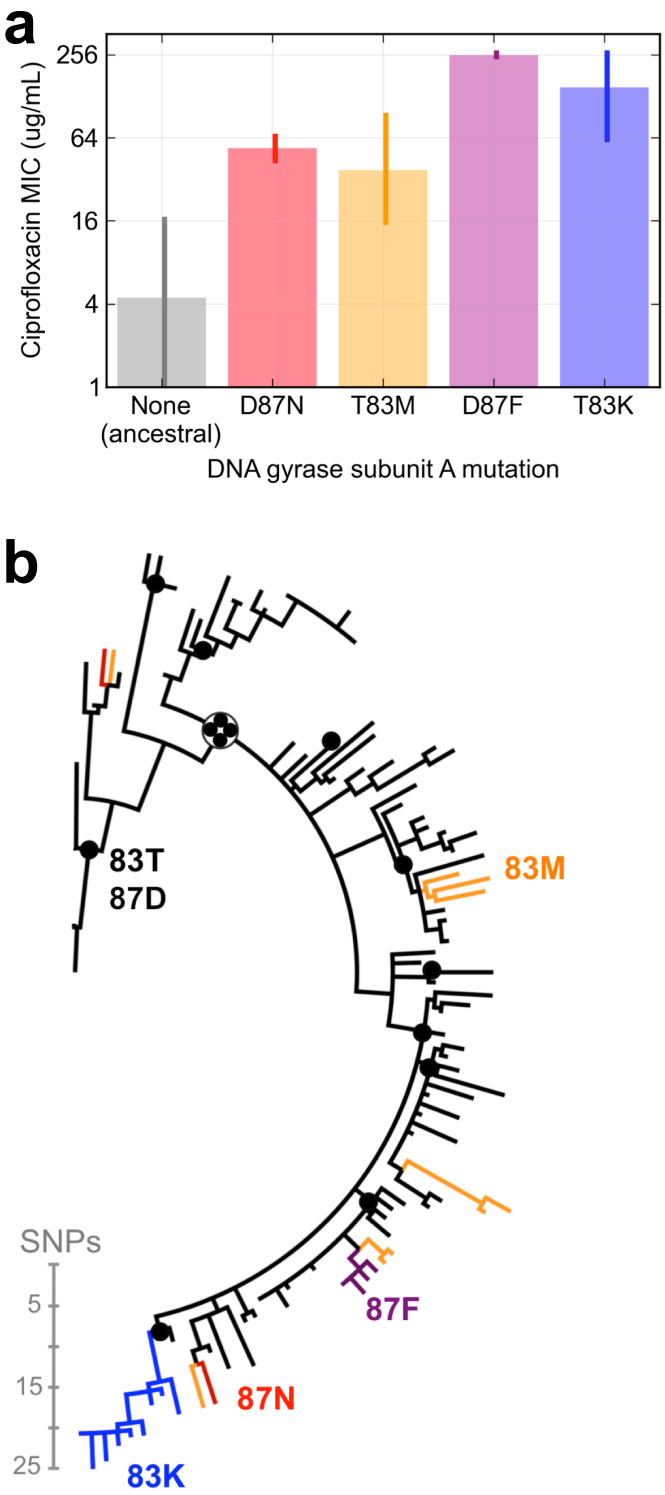

Figure S1.5. Resistance to ciprofloxacin as a function of mutations in BDAG_02180 (gyrA). a, We assayed the resistance to ciprofloxacin to all 112 isolates (precision: 2-fold). We plot on a logarithmic scale the average resistance of these isolates as a function of the mutation observed in BDAG_02180.Thin bars indicate the range of MIC observed for mutation. The number of isolates corresponding to each mutation is as follows: wildtype: 88, D87N: 2, T83M: 9, D87F: 4, T83K: 9. **b, The last common ancestors of each patient have drug-sensitive genotype.** We display the phylogeny of the 112 strains. Branches are colored as a function of the inferred genotype in the gene BDAG_02180 at codons 83 and 87: black corresponds to the wild type—a threonine (T) at position 83 and aspartate (D) at position 87 (D), blue indicates a lysine (K) at position 83, orange indicates an methionine (M) at position 83, red indicates a asparagine (N) at position 87, and purple indicates a phenylalanine (F) at position 87. The last common ancestors of strains from each patient are shown as dots; they are colored with their inferred BDAG_02180 genotype. Even though mutations occurred in 6 patients, all the last common ancestors bear the wild-type BDAG_02180, which is associated with gene sensitivity (Figure 3A). This indicates that the fluoroquinolone resistance evolves within patients, rather than being transmitted from patient-to-patient.

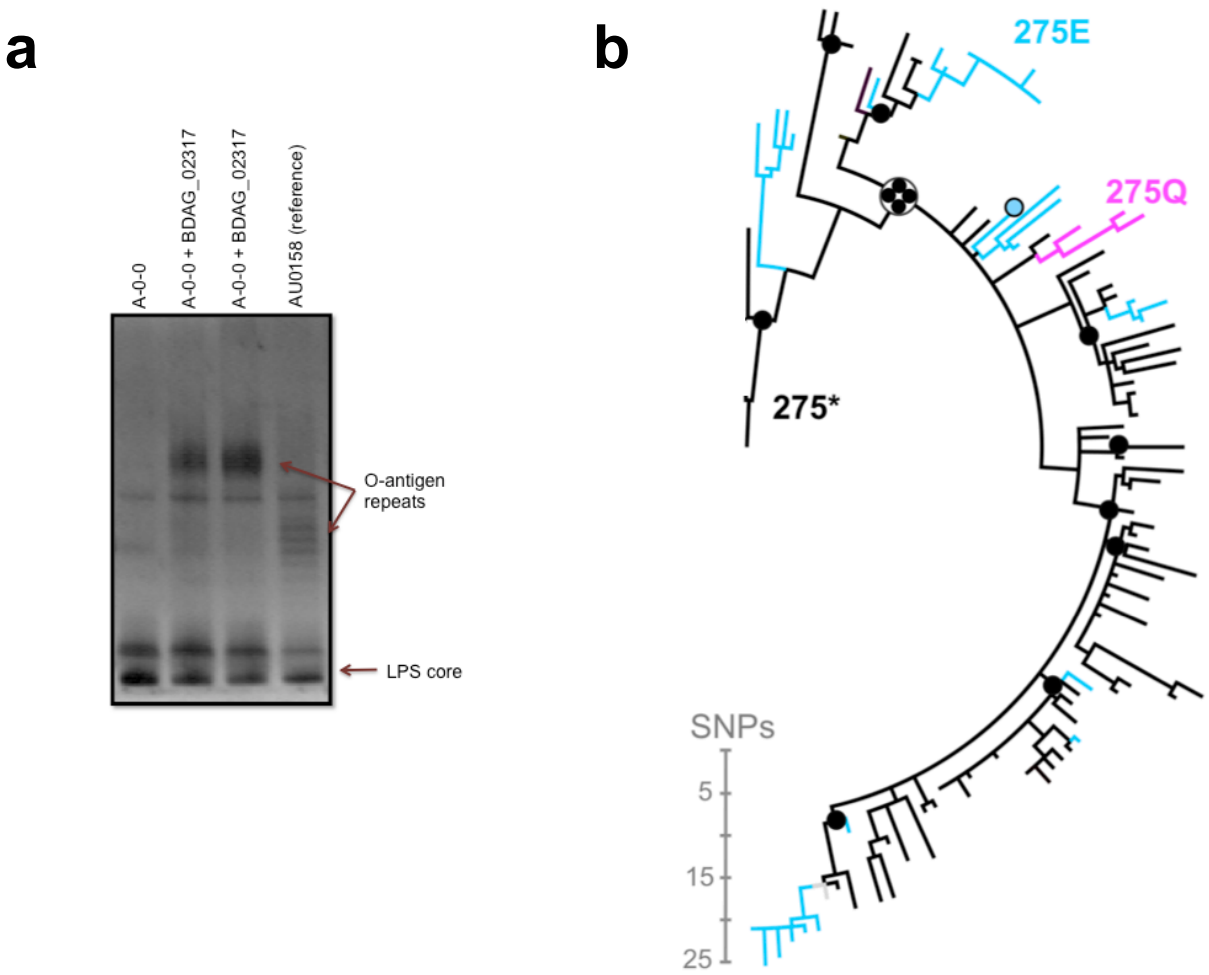

Figure S1.6. O-antigen repeats as a function of BDAG_02317. a, Complementation with a full-length BDAG_02317 restores O-antigen presentation. Full length glycosyltransferase BDAG_02317 was amplified from AU0158 (reference strain) and inserted into a plasmid under the control of a constitutive promoter (Supplementary Information 1). Transformation of strain A-0-0 with this plasmid restored O-antigen presentation. We note variation in O-antigen chain molecular weight, as we had also observed amongst our isolates (as seen in Figure 2.3b). **b, Most last common ancestors display a truncated glycosyltransferase.** We display the phylogeny of the 112 strains. Branches are colored as a function of the inferred genotype in the gene BDAG_02137 at codon 275 (black: stop codon, pink: glutamine , blue: glutamate, gray: unknown). The last common ancestors of strains from each patient are shown as dots; they are colored with their inferred BDAG_02137 genotype. Even though mutations occurred in 9 patients, all the last common ancestors but one (patient D, for whom we sequenced a single isolate) bear the stop codon. We hypothesize that the truncated genotype may provide an advantage during patient-to-patient transmission.

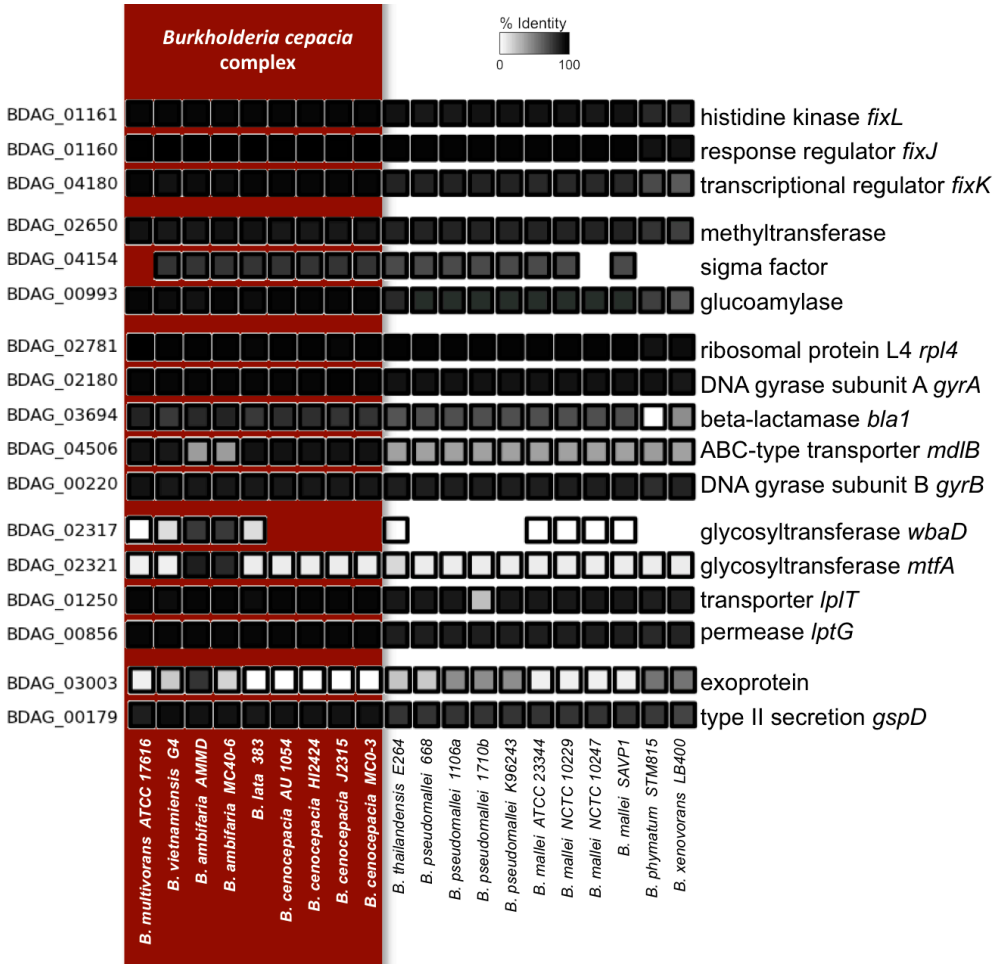

Figure S1.7: Most genes under positive selection are conserved across *Burkholderia*. For each gene under positive selection, blastp was used to find the nearest homologs in each of the 20 *Burkholderia* genomic sequences (9 of which are also members of the *Burkholderia cepacia* complex which colonize individuals with CF, highlighted in red). For each comparison, a box is drawn if a match was found with an E value of less than 10^{-8}, and that box is shaded in according to the percent amino-acid identity between the two sequences (see key). *B. mallei and B. pseudomallei* are important pathogens infecting healthy individuals.

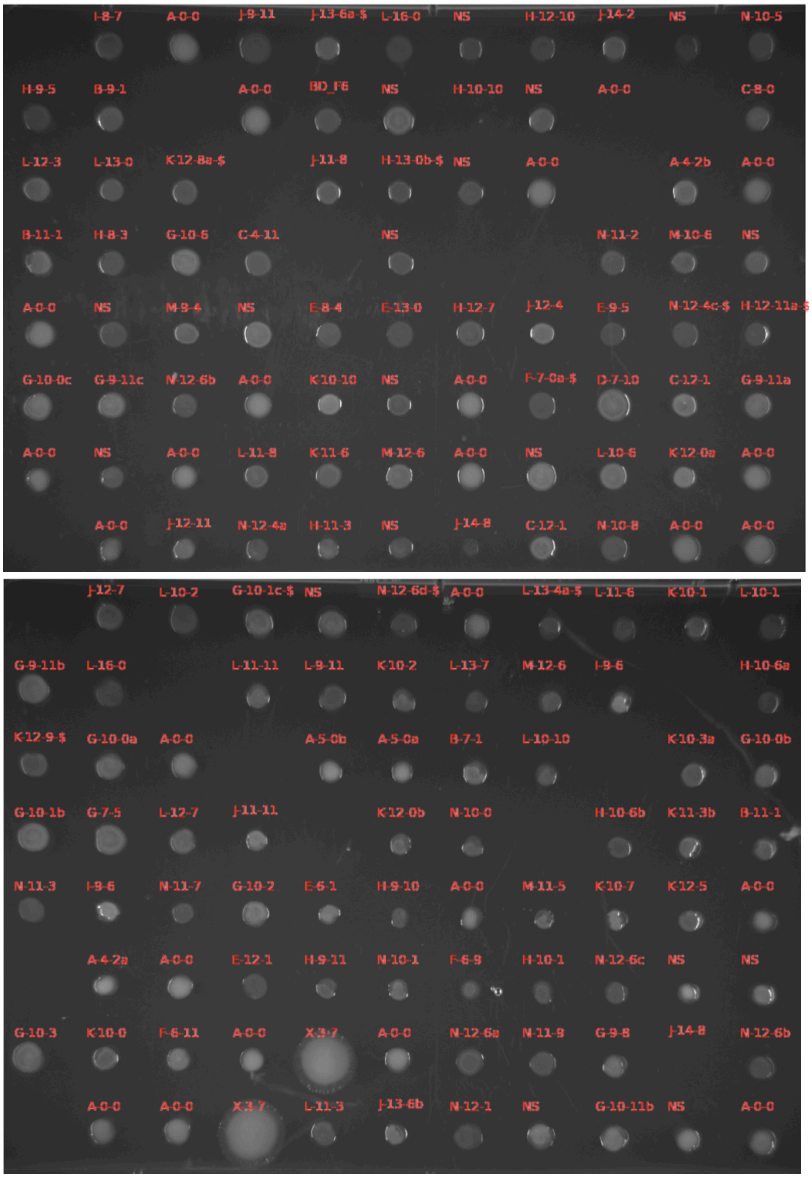

Figure S1.8. Isolates from the outbreak do not show significant variation in mucoidy. A fresh library of isolates was obtained by pinning from the thawed frozen stock into LB and growing overnight at 37°C. A 0.3 microliter aliquot of each overnight culture was pinned onto omnitrays containing YEM agar (0.5g Yeast Extract, 4 g mannitol, and 15g agar per liter) and grown at 37°C. After one day of growth, the plates were imaged. The outgroup strain, X-3-7, appears mucoid (right panel, two replicates). All strains from the outbreak are significantly less mucoid. Strains not sequenced for this study are indicated with “NS.” Not all sequenced strains are grown because some failed to grow on YEM agar

Table S1.1:
Burkholderia dolosa isolates used in this study

Study name	Sample origin	SRA Accession
Patient A		
A-0-0	Airway	SRS242477
A-4-2a	Airway	SRS242478
A-4-2b	Airway	SRS242497
A-5-0a	Airway	SRS242395
A-5-0b	Airway	SRS242479
Patient B		
B-4-11	Airway	SRS242400
B-7-1	Airway	SRS242414
B-9-1	Airway	SRS242401
B-11-1	Airway	SRS242412
Patient C		
C-4-11	Airway	SRS242415
C-8-0	Airway	SRS242484
C-10-0	Airway	SRS242416
C-12-1	Airway	SRS242485
Patient D		
D-7-10	Airway	SRS242441
Patient E		
E-6-1	Airway	SRS242402
E-8-4	Airway	SRS242417
E-9-5	Airway	SRS242428
E-12-1	Airway	SRS242430
E-13-0	Airway	SRS242480
Patient F		
F-6-9	Airway	SRS242481
F-6-11	Airway	SRS242482
F-7-0a-$	Blood	SRS242483
Patient G		
G-7-5	Airway	SRS242403
G-9-8	Airway	SRS242418
G-9-11a	Airway	SRS242419
G-9-11b	Airway	SRS242487
G-9-11c	Airway	SRS242420
G-10-0a	Airway	SRS242421
G-10-0b	Airway	SRS242422
G-10-0c	Airway	SRS242423
G-10-1b	Airway	SRS242424
G-10-2	Airway	SRS242488
G-10-1c-$	Blood	SRS242426
G-10-3	Airway	SRS242425
G-10-6	Airway	SRS242427
G-10-11b	Airway	SRS242404
Patient H		
H-8-3	Airway	SRS242386
H-9-5	Airway	SRS242410
H-9-10	Airway	SRS242493
H-9-11	Airway	SRS242442
H-10-1	Airway	SRS242443
H-10-6a	Airway	SRS242385
H-10-6b	Airway	SRS242444
H-10-10	Airway	SRS242445
H-11-3	Airway	SRS242446
H-12-7	Airway	SRS242447
H-12-10	Airway	SRS242472
H-12-11a-$	Blood	SRS242492
H-13-0b-$	Blood	SRS242471
Patient I		
I-8-7	Airway	SRS242411
I-9-6	Airway	SRS242413
Patient J		
J-9-11	Airway	SRS242466
J-11-8	Airway	SRS242439
J-11-11	Airway	SRS242406
J-12-4	Airway	SRS242440
J-12-7	Airway	SRS242407
J-13-6a-$	Blood	SRS242468
J-13-6b	Airway	SRS242467
J-12-11	Airway	SRS242469
J-14-8	Airway	SRS242408
J-14-2	Airway	SRS242470
Patient K		
K-9-0	Airway	SRS242463
K-10-0	Airway	SRS242489
K-10-1	Airway	SRS242432
K-10-2	Airway	SRS242433
K-10-3a	Airway	SRS242434
K-10-7	Airway	SRS242435
K-10-10	Airway	SRS242464
K-11-3a	Airway	SRS242431
K-11-3b	Airway	SRS242436
K-11-6	Airway	SRS242437
K-12-0a	Airway	SRS242405
K-12-0b	Airway	SRS242438
K-12-5	Airway	SRS242465
K-12-8a-$	Blood	SRS242490
K-12-9-$	Blood	SRS242491
Patient L		
L-9-11	Airway	SRS242473
L-10-1	Airway	SRS242448
L-10-2	Airway	SRS242449
L-10-6	Airway	SRS242409
L-10-10	Airway	SRS242387
L-11-3	Airway	SRS242388
L-11-6	Airway	SRS242389
L-11-8	Airway	SRS242390
L-11-11	Airway	SRS242391
L-12-3	Airway	SRS242392
L-12-7	Airway	SRS242393
L-13-0	Airway	SRS242394
L-13-4a-$	Blood	SRS242474
L-13-7	Airway	SRS242475
L-16-0	Airway	SRS242476
Patient M		
M-9-4	Airway	SRS242396
M-10-6	Airway	SRS242397
M-11-5	Airway	SRS242398
M-12-6	Airway	SRS242399
Patient N		
N-10-0	Airway	SRS242494
N-10-1	Airway	SRS242450
N-10-5	Airway	SRS242451
N-10-8	Airway	SRS242452
N-10-11	Airway	SRS242453
N-11-2	Airway	SRS242454
N-11-3	Airway	SRS242455
N-11-7	Airway	SRS242456
N-11-9	Airway	SRS242457
N-12-1	Airway	SRS242458
N-12-4a	Airway	SRS242459
N-12-4c-$	Blood	SRS242496
N-12-5-$	Blood	SRS242429
N-12-6a	Fluid, Right Inferior Hematoma	SRS242460
N-12-6b	Pleural fluid	SRS242461
N-12-6c	Tissue, Right Lung	SRS242462
N-12-6d-$	Blood	SRS242495
Patient X (outgroup)		
X-3-7	Airway	SRS242486

Study names designate patient and time. For example, 'A-2-4a' was recovered 4 years and 2 months after the date of the first isolate in our study, and it was the first isolate recovered from Patient A during this month. The second column designates origin of this isolate (blood isolates are also indicated in the study name by '$'. The third column indicates the Sequence Read Archive (at The National Center for Biotechnology Information, USA) sample number for the publicly available primary sequencing data

Table S1.2:
Full annotation and homology for genes under parallel, positive selection used to assess biological relevance.

Gene ID	Broad annotation	Number of mutations	Biological relevance	Annotated homologs: organism (query coverage)	Additional annotation
Genes not previously implicated in pathogenesis					
BDAG_01161	PAS	17	Oxygen-related gene regulation	*fixL*: *B. rhizoxinica* (98)	Two-component system histidine kinase
BDAG_01160	regulatory protein LuxR	4	Oxygen-related gene regulation	*fixJ*: *B. rhizoxinica* (99)	Two-component system response regulator
BDAG_04180	transcriptional regulator Crp/Fnr family	3	Oxygen-related gene regulation	*anr*: *B. multivorans* (99), *fixK*: *C. crescentus* (69)	
BDAG_02650	SAM-dependent methyltransferase	8	?	ZP_02908193.1: *B. ambifaria* (99)	Methyltransferase
BDAG_04154	DNA-directed RNA polymerase specialized sigma subunit	4	?	YP_001116246.1: *B. vietnamiensis* (86)	Gene regulation, ECF subfamily
BDAG_00993	Glucoamylase	3	?	ZP_02893540.1: *B. ambifaria* (99)	Glycoside hydrolase
Genes previously implicated in pathogenesis					
BDAG_02781	Ribosomal protein L4	14	Antibiotic resistance	*rpl4*	Macrolide resistance
BDAG_02180	DNA gyrase subunit A	11	Antibiotic resistance	*gyrA*	Quinolone resistance (see **Figure 3A**)
BDAG_03694	Beta-lactamase	8	Antibiotic resistance	YP_001811541.1: *B. ambifaria* (99), *bla1*: *B. cereus* (88)	Beta-lactam resistance
BDAG_04506	ABC-type multidrug transport system ATPase and permease components	5	Antibiotic resistance	*mdlB*: *M. succiniciproducens* (95)	Hypothetical role in AB resistance
BDAG_00220	Type IIA topoisomerase	3	Antibiotic resistance	*gyrB*	Quinolone resistance
BDAG_02317	Glycosyltransferase	10	Membrane Synthesis	*wbaD*: *E. coli* (88)	(see **Figure 3B**)
BDAG_02321	Glycosyltransferase	6	Membrane Synthesis	*mtfA*: *A. fulgidus* (99)	Mannitol transferase
BDAG_01250	Major facilitator superfamily (MFS_1) transporter	3	Membrane Synthesis	*lplT*: *B. cenocepacia* (99)	Lisophospholipid transporter
BDAG_00856	hypothetical protein	3	Membrane Synthesis	YP_002231781.1: *B. cenocepacia* (99), *lptG*: *E. clocae* (93)	Permease YjgP/YjgQ family
BDAG_03003	Large exoprotein involved in heme utilization or adhesion	4	Secretion	ZP_02889734: *B. ambifaria* (97), *fhaB*: *P. ananatis* (80)	Secreted product
BDAG_00179	Type II secretory pathway component PulD	3	Secretion	*gspD*: *B. cenocepacia* (99)	Secretion system component

Broad annotations are automatically assigned in the reference genome. Most relevant known homologs, additional annotation, and biological relevance were determined by manual annotation.

Table S1.3: Primers used in this study

Name	Primers 5'->3'
GlycTcompF	TTTTTGAGCTCATACGACGTCGATGCCGGAGATCG
GlycTcompR	TTTTTTCTAGACCGTCGCTCCGGAGTCTCAACC
pUCP18up	GGCTCGTATGTTGTGTGGAATTGTGAGCGG

Supplementary Information 1

Burkholderia dolosa

Burkholderia dolosa is a member of the *Burkholderia cepacia* complex*,* a group of related species capable of infecting immune-compromised hosts. Members of the *B. cepacia* complex were initially categorized as *Pseudomonas cepacia.* Members of this complex are found frequently in the soil, and are associated with the roots of plants (they were first identified as the cause of onion rot in 1950).

However, in the 1980s they were recognized as an emerging cause of infection in patients with cystic fibrosis[1]. Some members of the *B. cepacia* complex*,* including *B. dolosa,* cause deadly 'cepacia syndrome' – a rapid health decline characterized by fevers, necrotizing pneumonia and bacteremia (presence of bacteria in the bloodstream)[2].

B. dolosa is one of the rarest members of the *B. cepacia* complex. It was first established as a distinct species in 2004[3].

Patient cohort

We retrospectively studied a small epidemic of *Burkholderia dolosa* that affected 39 people with cystic fibrosis (CF) who were treated at a Boston hospital. The average age of these patients at date of first *B. dolosa* detection was 19 years.

Treatment of these patients for *B. dolosa* and other bacterial infections of CF patients includes frequent administration of antibiotics, usually in combination[4]*.* While antibiotics are often helpful in quelling an acute flare-up, they are generally ineffective in eradicating the infection. Of the 34 patients who continued care at the Boston hospital, only 3 eventually cleared the infection. Some of the patients still colonized remain asymptomatic, while others developed cepacia syndrome[5]. At the time of submission, 9 patients had received lung transplants; 17 patients had died, all but one from their illness.

We chose to study 14-deidentified patients, including patient zero of the epidemic. Seven patients were included in this study on the basis of availability of isolates from the blood; the remaining six were chosen at random among patients who did not have bacteremia. At the time of writing, 5 of these patients had received lung transplants; 8 patients had died, all but one from their illness. None of these patients cleared their airways of *Burkholderia dolosa*.

As mentioned above, these patients receive frequent antibiotic therapy. However, the patients' records do not allow us to know their history of antibiotic usage with high enough accuracy to correlate this data with antibiotic-resistance or data related to antibiotic-resistance.

Bacterial isolate collection

Samples from these patients were collected during normal care. The 97 isolates labeled "airways" were obtained from either sputum or bronchoalveolar lavage (BAL) fluid. Another 11 isolates labeled "blood" were sampled from the blood cultures. The remaining 4 isolates labeled "other" were sampled from pleural tissue pleural or mediastinal tissue from fluid obtained during surgical procedures. Details are in Table

S1.1. In most cases, the frozen clinical stocks were prepared from a single colony. The timing between isolates is described in Figure S1.1.

Frozen clinical stocks were streaked on solid media. A single colony from each plate was chosen at random and frozen in 15% glycerol to create a working library.

We also obtained the isolate LMG18943 (also known as AU0645), shown elsewhere to be of the same strain[6] (Strain SLC6). This isolate was taken from a different CF patient in a different location in the USA during the duration of the epidemic. We used this isolate as the outgroup for phylogenetic analyses.

Throughout this work, time of isolation is reported relative to the collection of an isolate from patient zero (isolate A-0-0). The time of collection of this isolate is unrelated to the first detection of *B. dolosa* in this patient. Letters distinguish isolates collected on the same month from the same patient. In some cases, the lettering is not continuous; some sequenced isolates were omitted from the study because they were either duplicates or recovered during autopsy.

Illumina sequencing

DNA was extracted from single colonies using standard procedures, following instructions from MoBio UltraClean® Microbial DNA Isolation Kit (cat # 12224-50, MO BIO Laboratories, Inc., USA).

Genomic libraries were constructed using the Illumina-compatible Nextera™ DNA Sample Prep Kit (cat# GA091120, EPICENTRE Biotechnologies, USA). Each genomic library was made up of four to six multiplexed genomes, barcoded with Nextera™ adapters. The libraries were sequenced using single-end, 75 bp reads on Illumina Genome Analyzer GAII by Partners HealthCare Center for Personalized Genetic Medicine (PCPGM). On average, 4.6 million reads were obtained per isolate (Figure S1.2a).

Reads were aligned using the *B. dolosa* genome AU0158 as a reference[7]; this reference genome belongs to an isolate recovered from patient zero. The reference genome is 6.42 megabases (Mb) long and comprises 3 chromosomes of lengths 3.41 Mb, 2.18 Mb, and .83 Mb. Alignment was performed by PCPGM using the standard Illumina pipeline[8] (CASAVA v1.6). We used SAMtools 0.1.12a to manipulate consensus sequences and find SNP between isolates[9].

For each strain we calculated the average number of reads aligned to each position in the reference genome (Figure S1.2b). We found the average read depth to be 37x. These reads produced high coverage: on average we obtained confident calls for 93% of the genome (Phred score>25).

Polymorphic loci between the 112 isolates

We first obtained the list of 4,375 loci where at least one of the 113 isolates (including the outgroup) was different than the reference genome. Next, we filtered out data to obtain a high-quality list of 511 polymorphic loci, where we expect less than one false positive.

Specifically, we used a two-step process. First, we retained loci where polymorphism was observed between at least two strains called with high confidence scores (Phred score >35). Then, at each of these 511 loci, we included all calls that were made with high enough confidence (Phred score >25; because fewer calls were made compared with the previous list, the Phred score could be lowered).

With this procedure, we obtain 56,811 calls at these 511 polymorphic loci. This two-step process and choice of scores maximizes the number of isolates with information per location, under the constraint of obtaining at most one miscalled base among all 56,811 calls. On average, 111 out of 113 isolates were called at these loci.

19 of these loci are unique to the outgroup; we resolve 492 polymorphic loci within the isolates of this epidemic.

Intragenic bias

We calculated that 426 of the 561 mutations occurred in intragenic regions (using the 5012 genes automatically annotated on the reference genome): the proportion of intragenic mutations is 75% (95% CI=72%-79%, Clopper-Pearson binomial confidence). This estimate is not statistically distinct from the intragenic proportion of the B. dolosa genome (79%): we do not detect a significant intragenic bias (p=0.04, binomial sampling with n=561).

Strains used for O-antigen assay

The strains used in the O-antigen assay presented in Figure 2.3b were (from left to right): A-0-0, AU0158, M-9-4, G-9-8, G-9-11b, C-12-1, B-11-1, E-9-5, E-12-1, C-10-0, K-10-0, K-12-5, K-12-8a-$, F-6-9, I-9-6, N-12-1, J-13-6a-$, J-14-2, J-14-8, and J-12-4.

Glycosyltransferase complementation

Escherichia coli Sm10[10] and *B. dolosa* were routinely grown on Luria agar (LA) containing 1.5% Bacto Agar (Difco, Franklin Lakes, N.J.). The medium was supplemented with 10 µg of tetracycline per ml for *E. coli* and 75 µg of tetracycline and 100µg of ampicillin per ml for *B. dolosa*, as appropriate. The glycosyltransferase gene BDAG_02317 was amplified by PCR with primers GlycTcompF and GlycTcompR (see Table S3.5) and genomic DNA of B. dolosa AU0158. This gene was ligated into the broad-host-range vector pUCP18Tc creating expression plasmid pUCP18Tc-GlycT. Biparental matings were performed to transfer pUCP18Tc-GlycT from E. coli SM10 to *B. dolosa* as previously described[11]. Transconjugants were confirmed by PCR using a forward primer specific of the complementing plasmid (pUCP18up) and primer GlycTcompR. O-antigen presentation was assayed as described in Online Methods.

fix genes

Burkholderia dolosa lacks the genes necessary for nitrogen fixation[1] suggesting that the *fix* system is utilized for gene-regulation of a different nature, as has been seen in other species[12]. All of the 24 mutations seen in this pathway are nonsynonymous and none result in a stop codon, suggesting a tuning of the pathway rather than its interruption. For this reason, we annotate these genes as involved in oxygen-related gene regulation.

Supplementary References

1. Coenye, T., Vandamme, P., *Burkholderia: molecular microbiology and genomics*. Horizon Scientific Press (2007).

2. Kalish, L. A. *et al.* Impact of *Burkholderia dolosa* on lung function and survival in cystic fibrosis. *Am J Respir Crit Care Med* **173**, 421-425 (2006).

3. Vermis, K. *et al.* Proposal to accommodate *Burkholderia cepacia* genomovar VI as *Burkholderia dolosa* sp. nov. *Int J Syst Evol Microbiol* **54**, 689-691 (2004).

4. Lyczak, J.B., C.L. Cannon, and G.B. Pier, Lung infections associated with cystic fibrosis. *Clinical microbiology reviews* **15**, 194-222 (2002).

5. Lipuma, J. J. The changing microbial epidemiology in cystic fibrosis. *Clin Microbiol Rev* **23**, 299-323 (2010).

6. Biddick, R., Spilker, T., Martin, A. & LiPuma, J. J. Evidence of transmission of *Burkholderia cepacia*, *Burkholderia multivorans* and *Burkholderia dolosa* among persons with cystic fibrosis. *FEMS Microbiol Lett* **228**, 57-62 (2003).

7. *Burkholderia dolosa* Sequencing Project. Broad Institute of Harvard and MIT (http://www.broadinstitute.org/).

8. Illumina, I., San Diego, CA., *Complete Secondary Analysis Workflow for the Genome Analyzer,* Pub. No. 770-2009-033 Current as of 19 October 2009.

9. Li, H., et al., The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-9 (2009).

10. Simon, R., P. U., and A. Puhler. A Broad Host Range Mobilization System for In Vivo Genetic Engineering: Transposon Mutagenesis in Gram Negative Bacteria. *Biotechnology* **1,** 784–91 (1983).

11. Urban, T. A. *et al.* Contribution of *Burkholderia cenocepacia* flagella to Infectivity and Inflammation. *Infection and Immunity* **72,** 5126-34 (2004).

12. Crosson, S., McGrath, P. T., Stephens, C., McAdams, H. H. & Shapiro, L. Conserved modular design of an oxygen sensory/signaling network with species- specific output. *Proc Natl Acad Sci U S A* **102**, 8018-8023 (2005).

Appendix 2:

Supplemental Materials for Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures (Chapter 3)

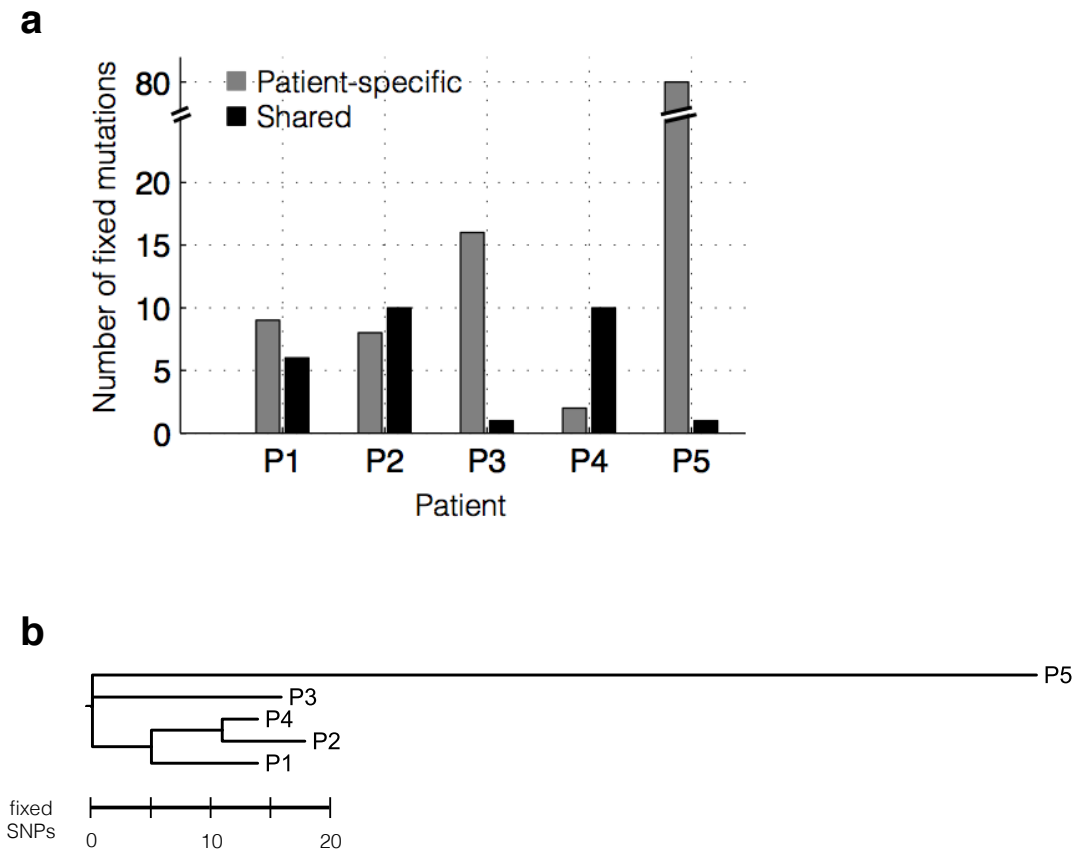

Figure S2.1: Some fixed mutations may have arisen and fixed prior to patient colonization. (**a**) Number of patient-specific and shared fixed mutations per patient. Shared fixed mutations are found to be fixed in some, but not all sputum samples, while patient-specific mutations are fixed in only one patient's sample. (**b**) The fixed mutations for each patient define a patient LCA, and we generate a maximum parsimony phylogeny among patient LCAs. The presence of interior branches in this phylogeny illustrates that some shared fixed mutations likely arose in an LCA of multiple patients. This tree was generated using the dnapars package in Phylip[1] and visualized in Figtree.

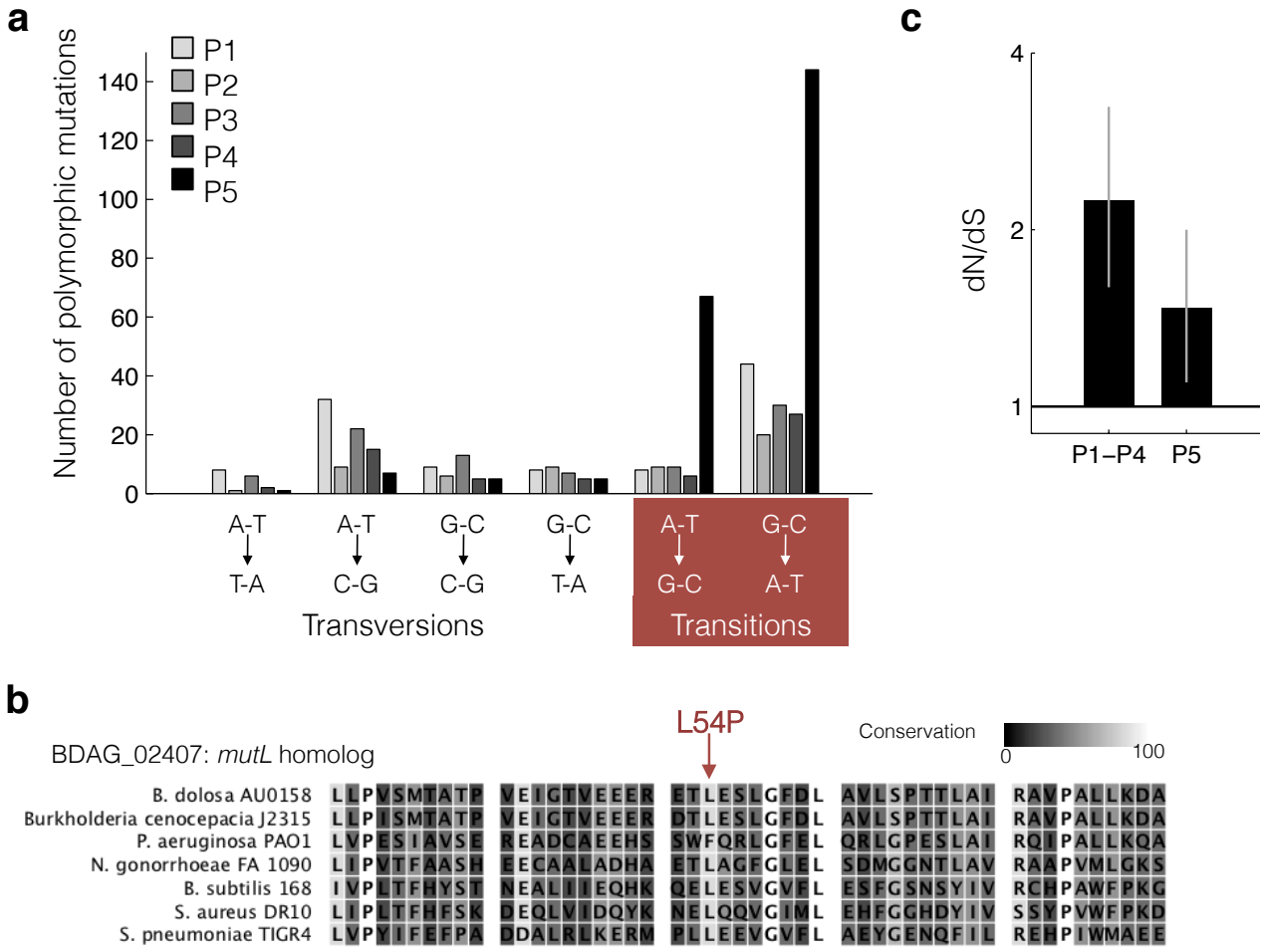

Figure S2.2: Excess mutations in Patient 5 are due to hypermutation. (**a**) Polymorphic mutations found within each patient's population were classified into 6 categories, without regard for strand (e.g. A->C is equivalent to T->G). P5 has an excess of both types of transition mutations (P<.001, Grubb's test for outliers) but not transversion (P<.001, Grubb's test for outliers) mutations, consistent with the known spectrum of mutations caused by *mutL* defects. (**b**) We scanned the annotations of the genes with point mutations in P5 for "DNA"' and manually inspected the results for roles in DNA repair. Only BDAG_02407 met this criteria. An NCBI BLAST search revealed this as a homolog of *mutL,* an essential component of mismatch repair. Other *mutL* sequences from NCBI gene were aligned and conservation was calculated using CLC Sequence Viewer 6. (**c)** P5's population has an increased relative rate of synonymous mutations. dN/dS was assessed as listed in the Methods for the intragenic mutations found in P1-P4 (non-hypermutators, n=278) and the intragenic mutations found in P5 (n=242). As described in the Methods, our approach for dNdS accounts for the decreased expected N/S in the hypermutators. Gray bars indicate 95% confidence interval. P < .001, one-side binomial z-test comparing the observed to the N/S expected under a dN/dS of 2.5 (P1-P4) and P5's spectrum of mutations.

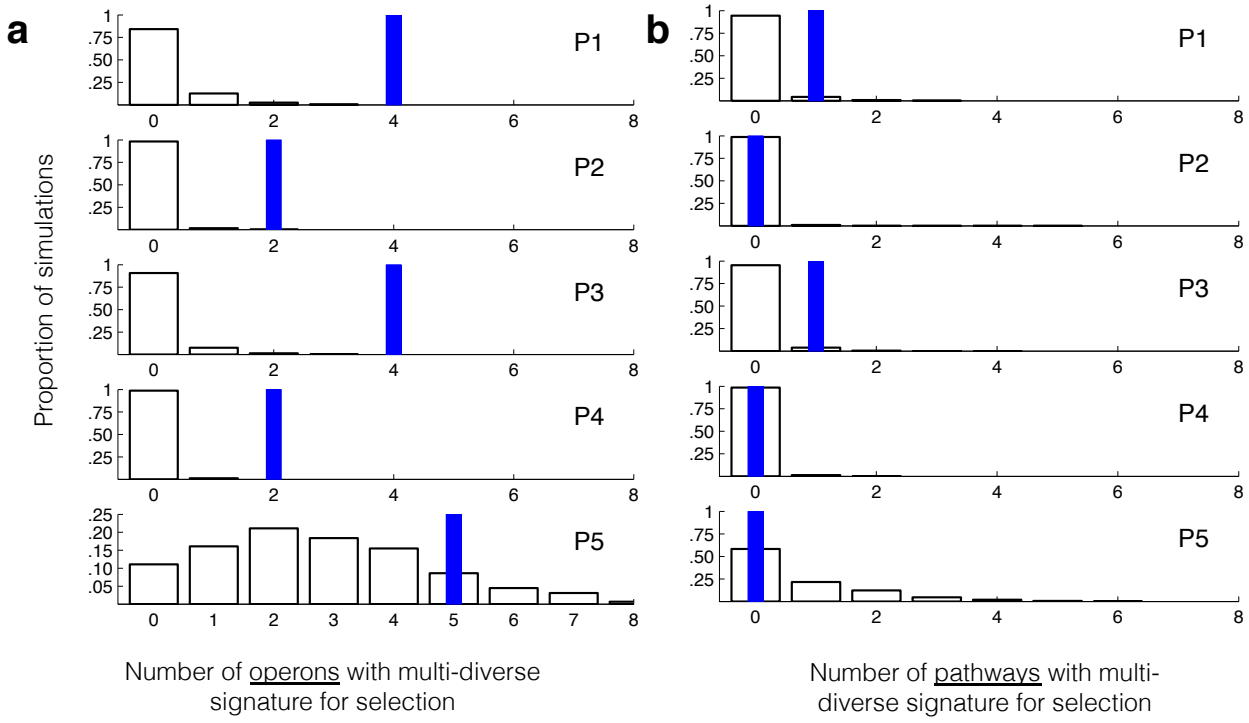

Figure S2.3: Search for operon and pathways undergoing parallel evolution within patients. We expanded our search for selective pressures to the operon and pathway levels, applying the same requirements of multiple mutations and more than one mutation per 2000 bp. (**a**) Multiple operons have this multi-diverse signature for selection in each patient (blue bars), while under neutrality we expect none in Patients 1-4 (P1-4). In P1-P4, the observed number of operons with multiple mutations is greater than the number obtained in > 995/1000 simulations (white bars show the results of 1000 simulations). For most of the 9 operons, the mutations that triggered their identification were all concentrated in a single, previously identified gene. The operons not previously implicated are shown in Table S2.3 (**b**) We did not find enrichment at the pathway level, suggesting that parallel evolution primarily acts at or below the gene level and supporting the idea that binning over larger regions of the genome can dilute the signal for selection. For each patient, we observe multiple mutations in the same pathway in over 18% of simulations (white bars). Relaxation of the requirement of mutations per bp in pathway brings up more multiply mutated genes in the simulations and not many more pathways. The two pathways multiply mutated per 2000 bp contain genes already found at the gene and operon level.

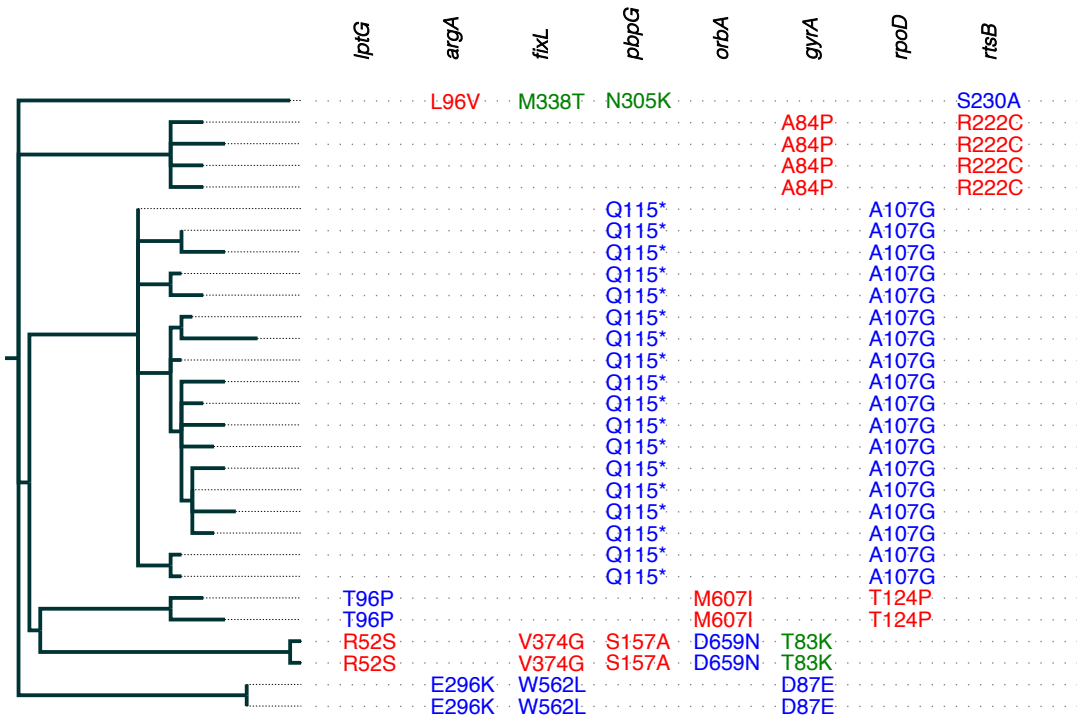

Figure S2.4: Colony re-sequencing from Patient 1 indicates within-patient parallel evolution. Seven genes within Patient 1 showed a multi-diverse signature for selection in the colony re-sequencing approach. The gene names are listed at top (see Tables S2.2 and S2.3) and the phylogeny of 29 colony isolates from Patient 1 is shown at left (same as Figure 2.3a). For each isolate, any mutations found in that gene are indicated on the corresponding horizontal line and column. No isolate has multiple mutations in the same gene.

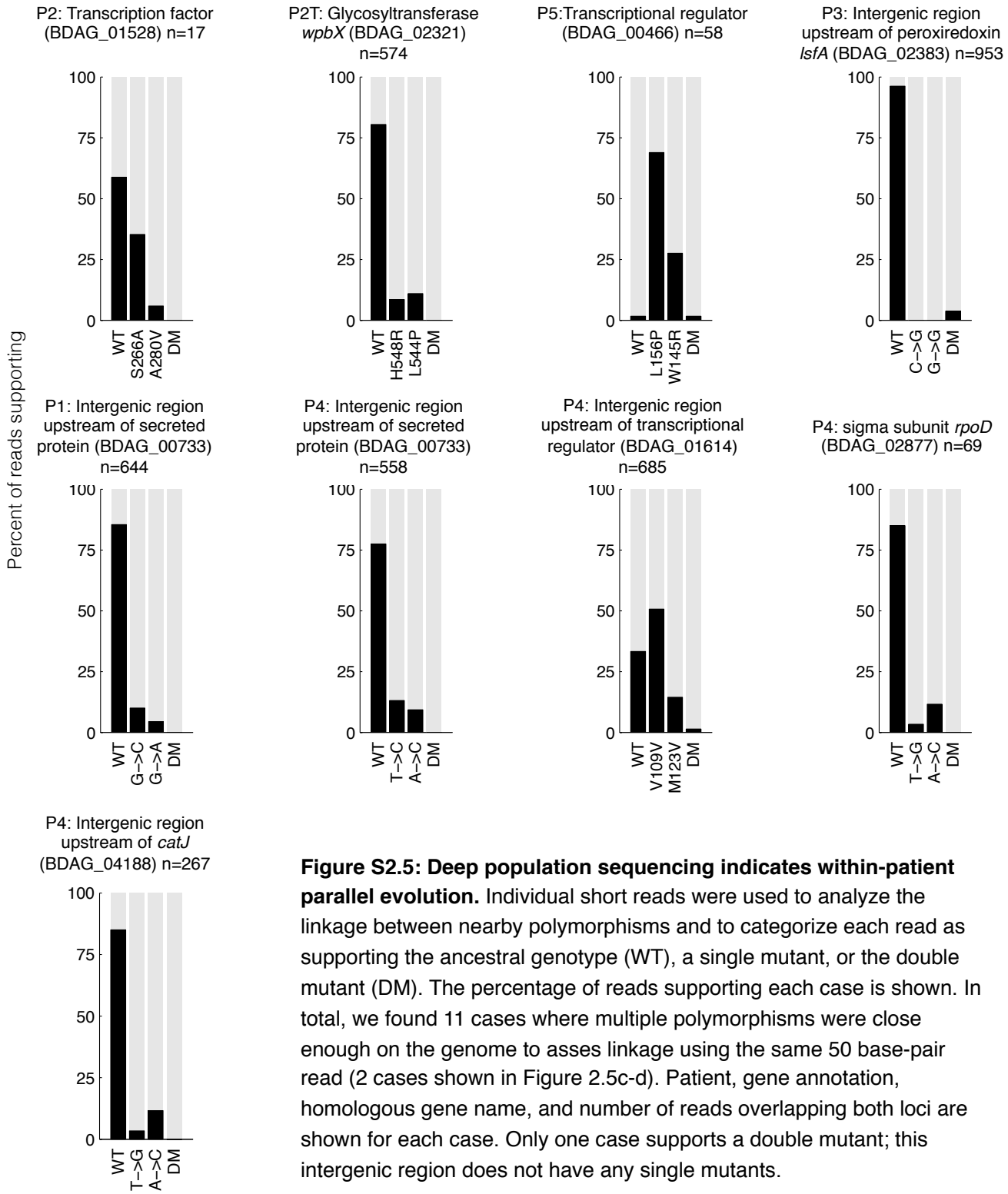

Figure S2.5: Deep population sequencing indicates within-patient parallel evolution. Individual short reads were used to analyze the linkage between nearby polymorphisms and to categorize each read as supporting the ancestral genotype (WT), a single mutant, or the double mutant (DM). The percentage of reads supporting each case is shown. In total, we found 11 cases where multiple polymorphisms were close enough on the genome to asses linkage using the same 50 base-pair read (2 cases shown in Figure 2.5c-d). Patient, gene annotation, homologous gene name, and number of reads overlapping both loci are shown for each case. Only one case supports a double mutant; this intergenic region does not have any single mutants.

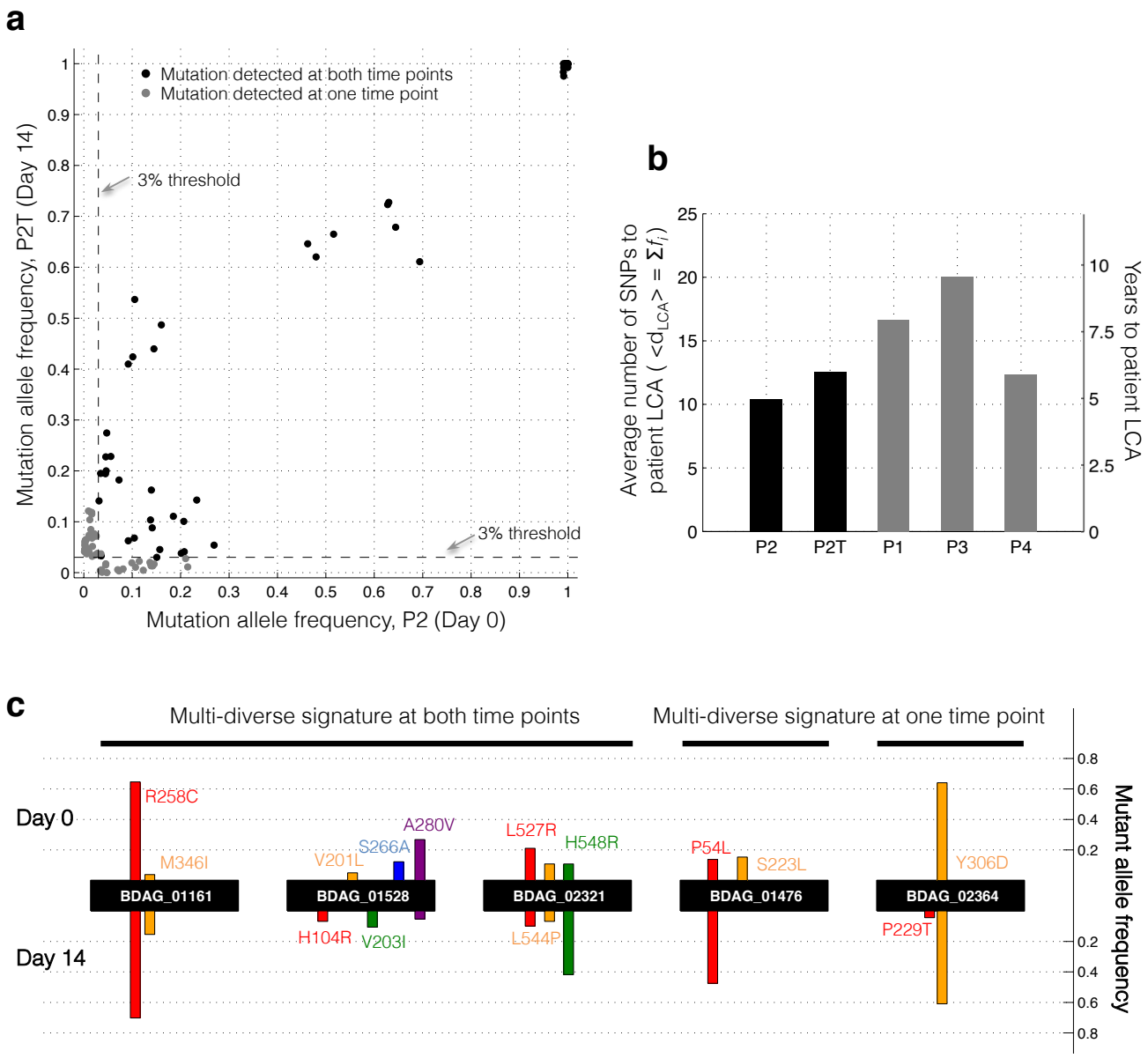

Figure S2.6: Comparison between two samples from same patient taken 14 days apart. (**a**) Scatter plot of mutant allele frequencies between samples shows agreement at very diverse positions with more variation in lower frequency mutations. (**b**) Both samples give similar estimates for time to patient LCA. See Supplementary Information 2 for discussion of error. (**c**) Three genes show a multi-diverse signature for selection in both samples, and each sample has only one gene displaying this signature that is not present in the other sample. Moreover, these two sample-specific genes show abundant mutations at both time points. See Methods for a description of the the two samples taken from this patient.

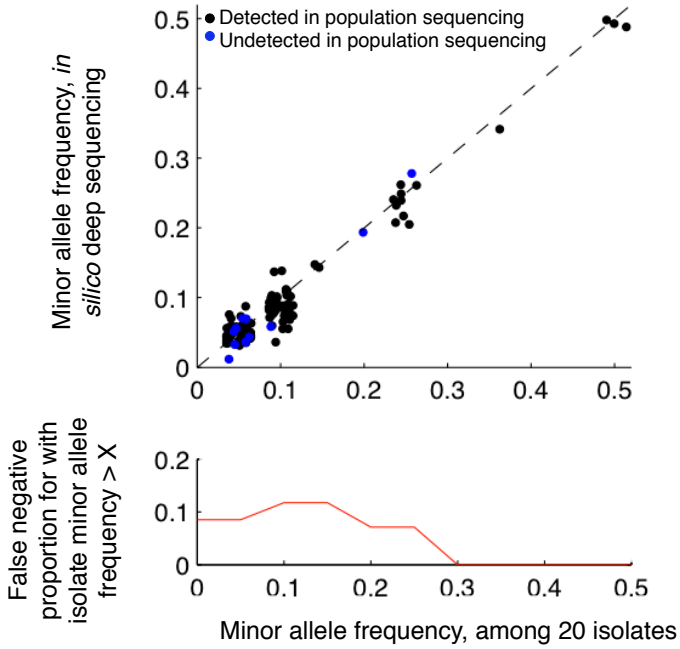

Figure S2.7: Deep population sequencing has low false positive rate and higher false negative rate. (a) We compared our approaches by mixing 20 isolates *in silico,* taking the same number reads from each of 20 isolates. We performed population deep sequencing analysis on this mixture of single-end reads, using the same isogenic control without paired-end information. All positives in the deep sequencing that were also found in the single colony isolates (no false positives). False negative positions (blue circles) fail the strand-bias filter or other quality filters (not shown). Jitter is added on the X-axis to improve visibility. **(b)** The proportion of positions called in the isolates not called in the *in silicio* deep sequencing (false negatives). 91% of positions called in the isolates were also called in deep sequencing (black circles).

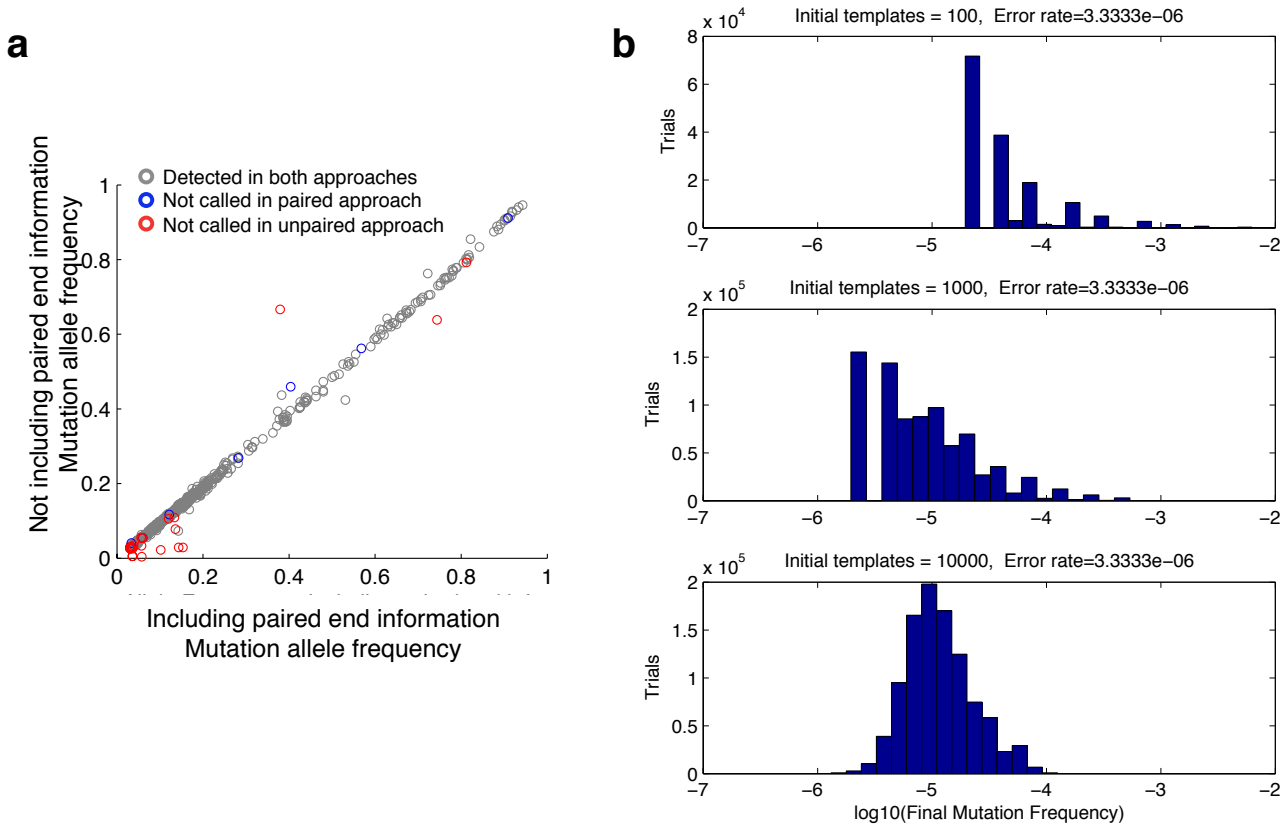

Figure S2.8: Paired-end information and Nextera amplification do not significantly affect polymorphism detection. (a) To investigate the effect of paired-end information on our results, we re-ran all population deep sequencing analyses, treating each read from a pair independently. For each genomic position, the mutation frequency from each approach is shown. Gray circles indicate positions called by both methods, blue circles indicate positions that did not pass all filters in the paired approach and red circles, and red circles indicate positions that did not pass all filters in the unpaired approach. 94% of positions called in the paired approach are also called in the unpaired approach and 98% of reads called in the paired approach are called in the paired approach. Bowtie2 sometimes fails to align unpaired reads well near short indels, and these positions have low coverage in the unpaired approach. Consistent with the fact that the paired approach discards pairs of reads when only one of the reads has poor sequencing quality (Illumina provided Phred scores), most of the discrepancy is attributable to noise around thresholds. The list of genes under selection within patients is identical using both approaches. **(b)** To understand the maximum possible frequency of errors introduced early during the 9-cycle PCR step of Nextera preparation, we performed simulations of PCR (described in Supplementary Information 2). Each simulation represents a unique single-nucleotide genomic position, and we ran 1 million simulations for each of 3 values of initial templates. We plot the histograms of final mutation frequency after 9 cycles for all genomic positions on a log scale (simulations with no mutations are not shown).

Table S2.1: Patient information at time of sample collection.

Patient	Years since acquisition based on clinical data	Sample Name	Analysis performed	FEV1 (% Predicted)	Approximate *B. dolosa* density in sputum sample (CFU/mL)*	Previous antibiotics (within 30 days)	Antibiotics at sample collection
1	7	P1	Colony re-sequencing (29 isolates) AND Deep population sequencing	57	4×10^7	None	None
2	8	P2	Deep population sequencing	57	6×10^7	Aztreonam (inhaled)	Aztreonam (inhaled)
		P2T (14 days after P2)	Deep population sequencing	58	4×10^8	Aztreonam (inhaled) Ceftazidime Minocycline Ciprofloxacin	Ceftazidime Minocycline Ciprofloxacin
3	9	P3	Deep population sequencing	61	4×10^7	Aztreonam (inhaled) Levofloxacin Minocycline Trimethoprim/Sulfamethoxazole Meropenem Ceftazidime	Aztreonam (inhaled) Levofloxacin Minocycline Chloramphenicol Meropenem
4	8	P4	Deep population sequencing	32	4×10^8	Levofloxacin Trimethoprim/Sulfamethoxazole Azithromycin	Azithromycin
5	9	P5	Deep population sequencing	57	5×10^8	Levofloxacin Minocycline Ceftazidime (inhaled) Azithromycin	Levofloxacin Minocycline Ceftazidime (inhaled) Azithromycin

*Colony forming units (CFU) per mL of frozen sample was calculated by serial dilution and plating.

Table S2.2: Genes with multi-diverse fingerprints for selection in one or more patients.

Predicted biological role	Gene number	Reference genome annotation	Annotated homolog [organism]*	Notes	CELLO[2] predicted localization	Patients mutated in [Mutations]**
Antibiotic resistance	*BDAG_ 01166*	D-alanyl-D-alanine carboxypeptidase	***pbpG*** [*Rubrivivax gelatinosus* IL144]	PBP7; Peptidoglycan biosynthesis; Beta-lactam resistance[3]; general stress[4].	Periplasmic	P1 [Q115*, S157A, N305K]
Antibiotic resistance	*BDAG_ 02180*	DNA gyrase subunit A	***gyrA*** [*Burkholderia gladioli* BSR3]	Fluroquinolone resistance; mutations in drug binding sites.	Cytoplasmic	P1 [A84P, T83K, T83M], P2 [*T83K*] P3 [*D87Y*], P4 [D87Y, T83K], P5 [*T83M*]
Outer membrane synthesis	*BDAG_ 00856*	hypothetical protein	***lptG*** *[Ralstonia sp.* PBA]	Permease YjgP/YjgQ family; LPS transport to the outer membrane[5].	Cytoplasmic Membrane	P1 [R52S, T96P], P2 [R268C] P3 [G323R], P4 [E63G, F306V]
Outer membrane synthesis	*BDAG_ 02321*	Glycosyltransferase	***wbpX*** [*Pseudomonas aeruginosa* PA7]	LPS synthesis[6].	Cytoplasmic	P2 [H548R, L527R]
Outer membrane synthesis	*BDAG_ 02311*	4-hydroxybenzoate polyprenyltransferase	***noeC*** [*Azorhizobium caulinodans* ORS 71]	D-arabinosylation; homologs *M. tuberculosis* and *P. aeruginosa* involved in cell-wall synthesis and pili glycosylation, respectively; adjacent to O-antigen biosynthesis genes here and in other organisms[7].	Cytoplasmic Membrane	P1 [W162*], P3 [A202A, S472R]
Iron scavenging	*BDAG_ 03997*	Outer membrane receptor protein	***huvA*** [*Vibrio anguillarum*]	Hemin transport system; Iron-regulated outer membrane heme receptor[8]; close homolog to several TonB-dependent heme receptors.	Outer membrane	P4 [V1M, L135R]
Iron scavenging	*BDAG_ 01606*	Outer membrane receptor protein	***orbA*** *[Burkholderia multivorans CF2]*	Siderophore receptor required for ferric ornibactin uptake[9]; mutations focused conserved barrel structure.	Outer membrane	P1 [M607I, D659N], P2 [*G547R*]
Lactate utilization	*BDAG_ 02124*	hypothetical protein	***lutC*** *[Bacillus licheniformis DSM 13]*	Iron-sulfur containing protein involved in lactate utilization[10]; also implicated in biofilm formation[11].	Cytoplasmic	P3 [A12E, E50A, G213G]
Oxygen-related gene regulation	*BDAG_ 01161*	PAS	***fixL*** [*Burkholderia rhizoxinica* HKI 454]	Two component system histidine kinase containing PAS domain, oxygen sensing[12]; also homologous to *P. aeruginosa bfiS*, biofilm development[13]; most mutations in or near heme binding pocket.	Cytoplasmic Membrane	P1 [M338T, V374G, W562L], P2 [R258C, M346I], P3 [*D445H*], P4 [M443R]
Unknown gene regulation	*BDAG_ 01528*	sigma54 specific transcriptional regulator, Fis family	YP_83538 [*Burkholderia cenocepacia* HI2424]	Transcriptional regulator containing PAS domain. Half of mutations focused in and near heme pocket, remaining near putative sigma54-interaction domain.	Cytoplasmic	P2 [V201L, S266A, A280V], P3 [L100R], P5 [A81V, H104R, E115A, A239T]
Unknown gene regulation	*BDAG_ 02877*	DNA-directed RNA polymerase sigma subunit	***rpoD*** [*Burkholderia cenocepacia* AU1054]	Homologous to primary sigma factor *rpoD*; mutated during experimental evolution of *B. cenocepacia*[14]*;* comparative analysis suggests *B. cepacia*[15] complex species have multiple recently evolved alternative primary sigma factors.	Cytoplasmic	P1 [A107G, T124P], P4 [V109V, M123V], P5 [*A87A, A95T*]+
Stringent Response	*BDAG_ 02219*	Guanosine polyphosphate pyrophosphohydrolase	***spoT*** [*Burkholderia ambifaria* AMMD]	(p)ppGpp metabolism; Role in virulence in *B. psuedomallei*[16]*.*	Cytoplasmic	P3 [S412L, I650L], P5 [R257H]
Arginine biosynthesis	*BDAG_ 01143*	Acetylglutamate kinase	***argA*** [*Cupriavidus necator* N-1]	Arginine biosynthesis.	Cytoplasmic	P1 [L96V, E296K], P5 [*R420H*]
Unknown	*BDAG_ 00993*	Glucoamylase	*ZP_02893540* [*Burkholderia ambifaria* IOP40-10]	Glycoside hydrolase. Trehalose synthesis.	Cytoplasmic	P4 [R77L, W276*, W299L, E403K]
Unknown	*BDAG_ 01476*	hypothetical protein	***rodZ*** [*Edwardsiella ictaluri* 93-146]	Homology to *E. coli rodZ* (component of bacterial cytoskeleton[17]) is weak and is an region that covers 1/3 of the gene.	Periplasmic	P2 [P54L, S123L]
Unknown	*BDAG_ 00061*	Hypothetical protein	YP_367612.1 [*Burkholderia sp.* 383]	Homologs in other *Burkholderia* but not many other genuses; no domains with known function.	Periplasmic	P3 [G95W, Q150*]

*Bolded gene names indicate that a reverse-BLAST suggests that the most homologous *B. dolosa* gene to the annotated gene is the queried gene (Methods). When database searches did not offer a candidate annotated homolog, the best BLAST hit is shown.

**Mutations are fixed in the patient's population if italicized, polymorphisms otherwise.

+PATRIC[18] suggests that this gene is likely misannotated, as homology to other *Burkholderia rpoD* genes starts at amino acid 92. There is a mutation prior to this, in Patient 5, which is likely noncoding but treated here as synonymous to remain systematic. Amino acids numbers here are relative to the genebank entry for BDAG_02877.

Table S2.3: Operons with multi-diverse fingerprints for selection in one or more patients.

Predicted biological role	Chromosome	Operon start*	Operon end*	Genes in operon mutated in at least one patient	Annotated homolog [organism]**	Patients mutated in [Mutations]***	Notes
Outer membrane synthesis	NZ_CH482380.1	729604	733200	*BDAG_00576*	***lptB*** [*Burkholderia sp.* KJ006]	P1 [D68A] P5 [*E243Q*]	LPS transport to the outer membrane[5].
				BDAG_00577	***lptA*** [*Burkholderia multivorans* ATCC 17616]	P1 [V70A]	
Outer membrane synthesis	NZ_CH482380.1	1052795	1055573	*BDAG_00856*	***lptG*** *[Ralstonia sp.* PBA]	P1 [R52S, T96P], P2 [R268C] P3 [G323R], P4 [E63G, F306V]	LPS transport to the outer membrane[5].
				BDAG_00857	***lptF*** [*Cupriavidus necator* N-1]	P3 [P289L], P4 [A154E]	
Unknown; implicated in various virulence-related pathways	NZ_CH482381.1	1240739	1242656	*BDAG_03702*	***rstB*** [*Serratia proteamaculans* 568]	P1 [R222C], P3 [R273P]	Two component histidine kinase and response regulator; *rtsAB* has been implicated as targets of the divalent cation sensing PhoQP system[19], though their targets and triggers are unknown: homologs of *rtsAB* are implicated in iron transport[20], biofilm-formation[21], degradation of virulence-related sigma factor RpoS[22], and acid shock and curli production[23].
				BDAG_03703	***rstA*** [*Burkholderia multivorans* ATCC 17616]	P3 [K108M], P4 [L176F]	

Operons that were detected exclusively on the basis of mutations focused in a single gene are not listed here and can be found in Table S2.2.

*Determined by FgenesB

**Bolded gene names indicate that a reverse-BLAST suggests that the most homologous *B. dolosa* gene to the annotated gene is the queried gene (Methods). When database searches did not offer a candidate annotated homolog, the best BLAST hit is shown.

***Mutations are fixed in the patient's population if italicized, polymorphisms otherwise.

Table S2.4: Coverage statistics

Sample	Type of reads	Percent of filtered reads aligned	Percent of reference genome callable*	Average coverage
Isogenic control	50bp paired end	94.8%	93.2%	708
P1	50bp paired end	96.3%	91.3%	582
P2	50bp paired end	95.3%	89.5%	489
P2T	50bp paired end	96.1%	90.7%	509
P3	50bp paired end	95.9%	89.8%	420
P4	50bp paired end	95.4%	89.8%	401
P5	50bp paired end	96.1%	89.1%	323
P1-01	50bp single end	96.5%	94.8%	36
P1-02	50bp single end	96.7%	94.7%	37
P1-03	50bp single end	96.7%	94.4%	34
P1-04	50bp single end	96.4%	94.7%	23
P1-05	50bp single end	96.5%	94.8%	37
P1-06	50bp single end	96.7%	93.3%	43
P1-07	50bp single end	96.9%	94.7%	51
P1-08	50bp single end	96.5%	94.8%	26
P1-09	50bp single end	96.9%	94.7%	40
P1-10	50bp single end	96.8%	94.4%	40
P1-11	50bp single end	96.8%	94.8%	60
P1-12	50bp single end	96.7%	94.8%	28
P1-13	50bp single end	96.6%	94.8%	36
P1-14	50bp single end	96.6%	94.4%	32
P1-15	50bp single end	96.6%	94.8%	29
P1-16	50bp single end	96.7%	94.7%	40
P1-17	50bp single end	96.8%	94.8%	48
P1-18	50bp single end	96.7%	94.7%	39
P1-19	50bp single end	96.6%	94.7%	25
P1-20	50bp single end	96.7%	94.8%	26
P1-21	50bp single end	96.5%	94.7%	34
P1-22	50bp single end	96.6%	94.7%	36
P1-23	50bp single end	96.8%	94.7%	48
P1-24	50bp single end	96.7%	93.4%	38
P1-25	50bp single end	97.1%	94.7%	39
P1-26	50bp single end	96.4%	94.6%	32
P1-27	50bp single end	96.9%	94.8%	37
P1-28	50bp single end	96.7%	94.8%	40
P1-29	50bp single end	96.4%	94.7%	25

*For isolates, callable positions are those with a consensus quality score (FQ, provided by samtools) score below -40. For population sequencing, callable positions are those that met coverage, base quality, mapping quality, and tail distance thresholds and for which the isogenic control had a major allele frequency of at least 98.5%

Supplementary Information 2

Sample collection

Expectorated sputum samples were collected at Boston Children's Hospital after written informed consent was obtained under protocols approved by the Institutional Review Boards at both Boston Children's Hospital and Harvard Medical School. Samples were immediately placed on ice and then liquefied with dithiothreitol. 10-15 mL phosphate buffered saline containing 1mM of dithiothreitol was added to each sample. Each sample was incubated on ice for 1 hour, vortexing every 20 minutes. 50% glycerol was added to each sample to a final concentration of 20%, and samples were then frozen at -80°C.

Sample prep, colony re-sequencing approach

Using a sterile 1-microliter plastic loop, ice was scraped from the top of the frozen homogenized sputum from Patient 1 and streaked onto the *Burkholderia cepacia* complex specific media, OFPBL (oxidation-fermentation basal medium supplemented with polymyxin B, bacitracin, lactose, and agar, BD diagnostic, USA). Individual colonies were picked into individual culture tubes containing 10mL of LB. Cultures were incubated with shaking at 37° C for 20 hours. Aliquots were stored in 15% glycerol at -80° C for further use. 1.8mL of this overnight culture was used for genomic DNA extraction.

Sample prep, deep population sequencing

Frozen sputum samples were thawed on ice. A 10-fold serial dilution was performed in PBS, and 0.8 mL of each dilution was plated on OFPBL using a disposable plastic spreader. Plates were dried and incubated at 37°C for 48 hours. Variation in colony size was observed within each sample. This variation may cause differences between allele frequencies measured and *in vivo* allele frequencies. Our results are robust to such differences; the signature for selection reported is insensitive to allele frequency (so long as mutations are above 3%) and all lineages, both those underrepresented and overrepresented, have been accumulating mutations since their LCA according to the molecular clock.

For each sample, a dilution plate was selected for harvesting which had between 5,000 and 30,000 small, mostly non-overlapping colonies. 2 mL of PBS was added and cells were scraped with a plastic loop and transferred to a microcentrifuge tube. A 0.5 mL aliquot of each sample was stored in 15% glycerol at -80°C for future use, and the remainder was used for DNA extraction. For the isogenic control, the same procedure was used, starting the serial dilution from a colony taken from Patient 1's sputum sample.

Initial data processing workflow for colony re-sequencing

We used custom MATLAB scripts to pipe together cutadapt[24] (remove adapter read-through), sickle[25] (trim low quality bases from reads), bowtie2[26] (align reads), and SAMtools[27] (call potential variants) and run them in parallel for many isolates on the Orchestra shared research cluster at Harvard Medical School. The following options were used:

> cutadapt -a CTGTCTCTTATACACATCTCTGA reads_1.fastq > trimmed_reads_1.fastq
> sickle se -f trimmed_reads_1.fastq -o filtered_reads_1.fastq -s singles.fastq -q 20 -l 25
> bowtie2 -X 2000 --no-mixed --very-sensitive --n-ceil 0,0.01 --un-conc unaligned.fastq -x refgenome_bowtie2 -U filteredreads.fastq -S aligned.sam
> samtools view -bS -o aligned.bam aligned.sam
> samtools sort aligned.bam aligned.sorted
> samtools mpileup -q30 -S -ugf refgenome.fasta aligned.sorted.bam > sample
> bcftools view -g sample > sample.vcf
> bcftools view -vS sample.vcf > variant.vcf

Mutations were called using custom MATLAB scripts and the vcf files.

Initial data processing workflow for deep population sequencing

We used custom MATLAB scripts to pipe together cutadapt, sickle, bowtie2, and SAMtools and run them in parallel for many isolates on the Orchestra shared research cluster at Harvard Medical School. The following options were used:

> cutadapt -a CTGTCTCTTATACACATCTCTGA reads_1.fastq > trimmed_reads_1.fastq
> cutadapt -a CTGTCTCTTATACACATCTCTGA reads_2.fastq > trimmed_reads_2.fastq
> sickle pe -f trimmed _reads_1.fastq -r trimmed _reads_2.fastq -o filtered_reads_1.fastq -p filtered_reads_2.fastq -s singles.fastq -q 20 -l 50
> bowtie2 -X 2000 --no-mixed --very-sensitive --n-ceil 0,0.01 --un-conc unaligned.fastq -x refgenome_bowtie2 -1 filteredreads_1.fastq -2 filteredreads_2.fastq -S aligned.sam
> samtools view -bS -o aligned.bam aligned.sam
> samtools sort aligned.bam aligned.sorted
> samtools mpileup -q30 -s -O -B -d3000 –f refgenome.fasta aligned.sorted.bam > sample.pileup
> samtools mpileup -q30 -S -ugf refgenome.fasta aligned.sorted.bam > sample
> bcftools view -g sample > sample.vcf
> bcftools view -vS sample.vcf > variant.vcf

The strict removal of all trimmed reads in sickle (-l 50 option) was used to make comparing tail distances across reads comparable. Custom MATLAB scripts were used to call diverse positions using the sample.pileup file. Fixed positions were called using custom MATLAB scripts and the vcf files.

Polymorphic mutation calling (deep population sequencing)

Using our isogenic control as a negative control, multiple isolates from Patient 1 as a positive control, and an interactive MATLAB environment that enabled investigation of the raw data, we developed a set of filters to identify polymorphic positions with minor allele frequency above 3%. We set the thresholds for these filters conservatively, minimizing false positives. An *in silico* mixing of reads from the isolates calls no positions not detected in the isolates and calls 78% of positions found when treating isolates separately (Figure S2.7). Similarly, 85.6% of positions called polymorphic above 3% frequency in Patient 1 using the pooled approach were also detected in the isolates (compared to 91.5% expected due to binomial sampling). No position is detected at greater than 13.5% frequency in population sequencing that is not detected in the isolates. All fixed and polymorphic mutations found and their frequencies are listed Table S2.6.

We considered a position to be polymorphic if it met the following quality thresholds in the given sample:

- **Minor allele frequency:** More than 3% of reads support a particular minor allele
- **Overall coverage:** At least 15 reads align in both the forward and reverse direction, and the total number of reads aligning is below the 99th percentile of covered positions in that sample.
- **Minor allele coverage:** At least 3 reads per major and minor allele aligning in both the forward and reverse direction (4 thresholds: forward minor, forward major, reverse minor, reverse major)
- **Base quality:** Average base quality (provided by sequencer) of greater than 19 for both the major and minor allele calls on both the forward and reverse strand
- **Mapping quality:** Average mapping quality (provided by aligner) of greater than 34 for reads supporting both the major and minor allele on both the forward and reverse strand
- **Tail distance:** Average tail distance of between 6 and 44 for reads supporting both the major and minor allele on both the forward and reverse strand
- **Indels:** Fewer than 20% of the reads aligning to that position support an indel at any position along that read
- **Strand bias:** A P-value of $> 10^{-5}$ supporting a null hypothesis that the minor allele frequency is the same for reads aligning to both the forward and reverse strand (Fisher's exact test).
- **Tail distance bias:** P-values of $> 10^{-5}$ supporting null hypotheses that the tail distances come from the same distribution for both the minor and major allele, for both the forward and reverse strand (t-test)

- **Isogenic control:** More than 98.5% of reads aligning to this genomic position in the isogenic control support a major allele

These filters remove false positive polymorphisms, which are not caused by randomly distributed sequencing error, but by systematic errors at particular genomic positions. For example, false positive polymorphisms can occur from misalignment near small insertions and deletions, from recent duplications represented once in the reference genome, and from neighboring sequences that increase the probability of sequencing error. Most systematic errors have hallmarks in individual reads data. Errors induced by a particular nearby DNA sequence, for example, will produce polymorphism on reads aligning in either the 5'->3' or 3'->5' direction, but not both. We do not consider regions with very high coverage, which may reflect recent or older duplications that are only listed once in the draft genome. We use the paired end information only to discard discordant read pairs and find that we would get similar results even without this information (see Figure 3.8a).

We find that false positive polymorphisms tend to be repeatable, showing nearly identical frequency and nucleotide identity across samples. Some genomic positions show repeatable polymorphism across samples and our isogenic control despite not having a hallmark for false-positive in the individual read data; such positions are removed by the requirement that the position be isogenic in the isogenic control (10-15 genomic positions per sample).

Estimation of false-positives from PCR amplification

While the Nextera kit involves PCR amplification, the large amount of template used and the limited number of cycles should prevent errors from PCR amplification from reaching near the 3% minor allele frequency threshold. We used two approaches to demonstrate this:

Simulation: To understand the maximum possible frequency of errors introduced early during the 9-cycle PCR step of Nextera preperation, we performed simulations of PCR. We conservatively assume that all PCR errors create the same mutated nucleotide, we assume perfect amplification, and we model a single genomic position (nucleotide) at a time. We assume a uniform error rate across the genome of $3.3 * 10^{-7}$ (provided by Epicentre). During each cycle of the simulation, new mutated molecules are introduced according to a Poisson process, the number of molecules doubles, and the number of previously mutated molecules doubles. We simulate 3 different values for the number of initial molecules covering each nucleotide. More initial molecules will buffer PCR errors. Given that the final concentration of DNA from the PCR reaction is 300ng and the genome size is 6.4 Mb, even coverage and perfect amplification predicts that there are 8.7×10^4 copies of each nucleotide in the initial pool (following tagmentation). Because genomic positions are certainly not represented evenly in this initial pool, we ran sets of simulation simulations using 10^4, 10^3, and 10^2 initial molecules (regions with lower abundances will not meet coverage requirements in final library). For each number of initial molecules, we simulate 10^6 nucleotide positions, a number larger than the number genomic positions with low representation in the initial pool. We find the maximum final frequency of mutation in 10^6 simulations with 10^2 initial molecules is .1%, much below our detection threshold of 3%. Simulations with more starting molecules have even lower error. See Figure S2.9b.

Empirical analysis of isogenic control: We have also empirically estimated less than 1 false positive polymorphic genomic position per sample introduced by PCR error per sample. In the isogenic control, positions introduced by PCR error will not be filtered out by the quality filters described above (all filters except the isogenic control filter). Thus, if PCR introduces false polymorphisms, we should see them in the isogenic control. 10 genomic positions in the isogenic control pass all quality filters. Of these 10, 9 show consistent polymorphism identity and frequency across all, indicating that when PCR error emerges, reproducibility of this PCR error allows it to be filtered out by the isogenic control.

Genetic distance to LCA calculation (deep population sequencing)

Here, we show that the average number of mutations per cell in a population is equivalent to the sum of the mutation frequencies in that population. The number of mutations in a given cell can be represented as $\sum_0^P m_i$, where P is the number of positions on the genome, and $m_i = 1$ if a cell as a mutation at position i and $m_i = 0$ otherwise. Thus:

$$\langle \text{Number of muations per cell} \rangle >= \langle \sum_{i=1}^{P} m_i \rangle$$

Now, we consider a population of C cells. The presence of a mutation at position i in a cell j is now indicated by $m_{ij} = 1$.

$$\langle \sum_{i=1}^{P} m_i \rangle = \frac{\sum_{j=1}^{C} \sum_{i=1}^{P} m_{ij}}{C} = \frac{\sum_{i=1}^{P} \sum_{j=1}^{C} m_{ij}}{C} = \sum_{i=1}^{P} \frac{\sum_{j=1}^{C} m_{ij}}{C} = \sum_{i=1}^{P} f_i$$

Where f_i is the mutation frequency at position i on the genome.

Error sources in estimation of time to LCA

Potential sources of error in estimating the time to LCA include Poisson error in the number of mutations accumulated in each lineage since the LCA, underestimation due to limited sensitivity in detecting mutations, overestimation due to false positives, and underestimation due to incomplete sampling of the diversity in the lung. Additionally, possible errors in converting $<d_{LCA}>$ to years include potential overestimation of the molecular clock resulting in underestimation of time to LCA (if historical sampling in the clinic is biased towards colonies with faster growth rates) and deviations from clock-like evolution (e.g. extra doublings following antibiotic treatment). The confidence intervals of time to LCA presented for the colony re-sequencing approach are calculated according to a Poisson distribution. We assume that the number of mutations in each cell is drawn from a Poisson distribution with λ equal to the mean across cells. Poisson measurement error is minimal for the population sequencing, as we are sampling hundreds of lineages simultaneously. The differences in estimated time to LCA between the two samples taken from Patient 2 are larger than would be expected given only Poisson error, suggesting other factors contribute to our ability to date the LCA (Figure S2.6b).

Gene annotation

Genes were annotated using a suite of online bioinformatics tools. Open reading frames were compared to RefSeq using NCBI's BLASTp and close homologs were scanned for gene names. For genes for which this did not indicate an obvious homolog, the Burkholderia Genome Database[28] and Microbial Genome Database for Comparative Analysis[29] (MBGD) were used to probe for candidate ortholog gene names and look for synteny (As *B. dolosa* is not in MBGD, homologs from other members of the *B. cepacia* complex were used to query MBGD).

When these searches produced candidate orthologs, reverse BLAST was used to test the orthology[30]; an ORF with that putative ortholog name from *E. coli* or other well-described non-*Burkholderia* genome was located in the NCBI gene database and used as a query for Blastp against the *B. dolosa* genome. If the original query came up as the best match and had an E value of < .001, this ortholog name was listed in Figure 3.5e. Otherwise, the number of the locus tag was listed. Biological relevance was assigned using literature search and various databases, including WikiGenes[31], STRING[32], UniProt[33], and PATRIC[18]. Subcellular localization was predicted using CELLO[2]. Literature searches were conducted to determine functional role of genes. Summary and gene-specific references can be found in Tables S2.2-3. Visualization for Figure 3.5e was performed using Cytoscape[34].

Supplementary References

1 Felsenstein, J. PHYLIP-Phylogeny Inference Package (Version 3.2). *Cladistics* **5**, 164-166 (1989).

2 Yu, C. S., Chen, Y. C., Lu, C. H. & Hwang, J. K. Prediction of protein subcellular localization. *Proteins* **64**, 643-651, doi:10.1002/prot.21018 (2006).

3 Gomez, M. J. & Neyfakh, A. A. Genes involved in intrinsic antibiotic resistance of Acinetobacter baylyi. *Antimicrobial agents and chemotherapy* **50**, 3562-3567, doi:10.1128/AAC.00579-06 (2006).

4 Kenyon, W. J. *et al.* Sigma(s)-Dependent carbon-starvation induction of pbpG (PBP 7) is required for the starvation-stress response in *Salmonella enterica* serovar *Typhimurium*. *Microbiology* **153**, 2148-2158, doi:10.1099/mic.0.2007/005199-0 (2007).

5 Ruiz, N., Kahne, D. & Silhavy, T. J. Transport of lipopolysaccharide across the cell envelope: the long road of discovery. *Nature Reviews Microbiology* **7**, 677-683 (2009).

6 Rocchetta, H. L., Burrows, L. L., Pacan, J. C. & Lam, J. S. Three rhamnosyltransferases responsible for assembly of the A-band D-rhamnan polysaccharide in *Pseudomonas aeruginosa*: a fourth transferase, WbpL, is required for the initiation of both A-band and B-band lipopolysaccharide synthesis. *Molecular microbiology* **28**, 1103-1119 (1998).

7 Harvey, H., Kus, J. V., Tessier, L., Kelly, J. & Burrows, L. L. *Pseudomonas aeruginosa* D-arabinofuranose biosynthetic pathway and its role in type IV pilus assembly. *The Journal of biological chemistry* **286**, 28128-28137, doi:10.1074/jbc.M111.255794 (2011).

8 Mourino, S., Rodriguez-Ares, I., Osorio, C. R. & Lemos, M. L. Genetic variability of the heme uptake system among different strains of the fish pathogen *Vibrio anguillarum*: identification of a new heme receptor. *Applied and environmental microbiology* **71**, 8434-8441, doi:10.1128/AEM.71.12.8434-8441.2005 (2005).

9 Sokol, P. A., Darling, P., Lewenza, S., Corbett, C. R. & Kooi, C. D. Identification of a siderophore receptor required for ferric ornibactin uptake in *Burkholderia cepacia*. *Infection and immunity* **68**, 6554-6560 (2000).

10 Smaldone, G. T., Antelmann, H., Gaballa, A. & Helmann, J. D. The FsrA sRNA and FbpB protein mediate the iron-dependent induction of the *Bacillus subtilis* lutABC iron-sulfur-containing oxidases. *Journal of bacteriology* **194**, 2586-2593, doi:10.1128/JB.05567-11 (2012).

11 Snitkin, E. S. *et al.* Tracking a Hospital Outbreak of Carbapenem-Resistant *Klebsiella pneumoniae* with Whole-Genome Sequencing. *Science Translational Medicine* **4**, 148ra116-148ra116 (2012).

12 Crosson, S., McGrath, P. T., Stephens, C., McAdams, H. H. & Shapiro, L. Conserved modular design of an oxygen sensory/signaling network with species-specific output. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 8018-8023, doi:10.1073/pnas.0503022102 (2005).

13 Petrova, O. E. & Sauer, K. A novel signaling network essential for regulating *Pseudomonas aeruginosa* biofilm development. *PLoS pathogens* **5**, e1000668, doi:10.1371/journal.ppat.1000668 (2009).

14 Coutinho, C. P., de Carvalho, C. C., Madeira, A., Pinto-de-Oliveira, A. & Sá-Correia, I. *Burkholderia cenocepacia* phenotypic clonal variation during a 3.5-year colonization in the lungs of a cystic fibrosis patient. *Infection and immunity* **79**, 2950-2960 (2011).

15 Menard, A., de Los Santos, P. E., Graindorge, A. & Cournoyer, B. Architecture of *Burkholderia cepacia* complex sigma70 gene family: evidence of alternative primary and clade-specific factors, and genomic instability. *BMC genomics* **8**, 308, doi:10.1186/1471-2164-8-308 (2007).

16 Muller, C. M. *et al.* Role of RelA and SpoT in *Burkholderia pseudomallei* virulence and immunity. *Infection and immunity* **80**, 3247-3255, doi:10.1128/IAI.00178-12 (2012).

17 Alyahya, S. A. *et al.* RodZ, a component of the bacterial core morphogenic apparatus. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 1239-1244, doi:10.1073/pnas.0810794106 (2009).

18 Gillespie, J. J. *et al.* PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infection and immunity* **79**, 4286-4298, doi:10.1128/IAI.00207-11 (2011).

19 Eguchi, Y. *et al.* Signal transduction cascade between EvgA/EvgS and PhoP/PhoQ two-component systems of *Escherichia coli*. *Journal of bacteriology* **186**, 3006-3014 (2004).

20 Jeon, J. *et al.* RstA-promoted expression of the ferrous iron transporter FeoB under iron-replete conditions enhances Fur activity in *Salmonella enterica*. *Journal of bacteriology* **190**, 7326-7334 (2008).

21 Bilecen, K. & Yildiz, F. H. Identification of a calcium-controlled negative regulatory system affecting *Vibrio cholerae* biofilm formation. *Environmental microbiology* **11**, 2015-2029 (2009).

22 Cabeza, M. L., Aguirre, A., Soncini, F. C. & Véscovi, E. G. Induction of RpoS degradation by the two-component system regulator RstA in *Salmonella enterica*. *Journal of bacteriology* **189**, 7335-7342 (2007).

23 Ogasawara, H. *et al.* Genomic SELEX search for target promoters under the control of the PhoQP-RstBA signal relay cascade. *Journal of bacteriology* **189**, 4791-4799 (2007).

24 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, pp. 10-12 (2011).

25 Sickle (https://github.com/ucdavis-bioinformatics/sickle).

26 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357-359 (2012).

27 Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

28 Winsor, G. L. *et al.* The Burkholderia Genome Database: facilitating flexible queries and comparative analyses. *Bioinformatics* **24**, 2803-2804, doi:10.1093/bioinformatics/btn524 (2008).

29 Uchiyama, I., Higuchi, T. & Kawai, M. MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic acids research* **38**, D361-365, doi:10.1093/nar/gkp948 (2010).

30 Wolf, Y. I. & Koonin, E. V. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome biology and evolution* **4**, 1286-1294, doi:10.1093/gbe/evs100 (2012).

31 Hoffmann, R. A wiki for the life sciences where authorship matters. *Nature genetics* **40**, 1047-1051, doi:10.1038/ng.f.217 (2008).

32 Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* **39**, D561-568, doi:10.1093/nar/gkq973 (2011).

33 UniProt, C. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research* **40**, D71-75, doi:10.1093/nar/gkr981 (2012).

34 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).